
S E
M S M
E S

Proceedings of the

# International Congress of Mathematicians

Madrid, August 22–30, 2006

## VOLUME III

# Invited Lectures

Marta Sanz-Solé
Javier Soria
Juan Luis Varona
Joan Verdera

Editors

European Mathematical Society

Editors:

Marta Sanz-Solé
Facultat de Matemàtiques
Universitat de Barcelona
Gran Via 585
08007 Barcelona
Spain

Javier Soria
Departament de Matemàtica Aplicada i Anàlisi
Facultat de Matemàtiques
Universitat de Barcelona
Gran Via 585
08007 Barcelona
Spain

Juan Luis Varona
Departamento de Matemáticas y Computación
Universidad de La Rioja
Edificio J. L. Vives
Calle Luis de Ulloa s/n
26004 Logroño
Spain

Joan Verdera
Departament de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona)
Spain

2000 Mathematics Subject Classification: 00Bxx

ISBN 978-3-03719-022-7

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available in the Internet at http://dnb.ddb.de.



Contact address:

European Mathematical Society Publishing House
Seminar for Applied Mathematics
ETH-Zentrum FLI C4
CH-8092 Zürich
Switzerland

Phone: +41 (0)44 632 34 36
Email: info@ems-ph.org
Homepage: www.ems-ph.org

Typeset using the author's TEX files: I. Zimmermann, Freiburg
Printed in Germany

9 8 7 6 5 4 3 2 1

# Contents

## 10 Ordinary differential equations and dynamical systems (continued)



## 11 Partial differential equations



*In case of several authors, invited speakers are marked with an asterisk.

## 12 Mathematical physics



## 13 Probability and statistics

## 14 Combinatorics

## 15 Mathematical aspects of computer science



## 16 Numerical analysis and scientific computing



## 17 Control theory and optimization

## 18 Application of mathematics in the sciences



## 19 Mathematics education and popularization of mathematics

## 20 History of mathematics



## Special activity

# Braids and differential equations

Robert Ghrist*

**Abstract.** Forcing theorems based on topological features of invariant sets have played a fundamental role in dynamics and differential equations. This talk focuses on the recent work of Vandervorst, Van den Berg, and the author using braids to construct a forcing theory for scalar parabolic PDEs, second-order Lagrangian ODEs, and one-dimensional lattice dynamics.

**Mathematics Subject Classification (2000).** Primary 37B30, 35K90; Secondary 34C25, 37L60, 57M25.

**Keywords.** Braids, Conley index, dynamical systems, parabolic PDEs, second order Lagrangian.

This talk covers a particular type of forcing theory for parabolic dynamics which uses the topology of braids in an index theory.

## 1. Topological forcing theorems

Throughout the last century of work in dynamical systems, forcing theorems have played a substantial role in establishing coarse minimal conditions for complicated dynamics. Forcing theorems in dynamics tend to take the following form: given a dynamical system of a specified class, the existence of some invariant set of one topological type implies the existence of invariant sets of other topological types. This forcing is often encoded by some sort of ordering on topological types of invariant sets.

### 1.1. Examples.

Three canonical examples of forcing theorems frame our work.

**Example 1** (*Morse Theory* [43]). The class of systems is that of nondegenerate gradient flows on an oriented manifold $M$. The invariant sets of interest are the fixed points, and the topological type associated to a fixed point is its *Morse index* – the dimension of its unstable manifold. A suitable chain complex generated by fixed points and graded by the Morse index yields a homology which is isomorphic to that

*Research supported by the National Science Foundation, PECASE DMS-0337713. The author wishes to thank Rob Vandervorst, without whom the work described here would not exist.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of $M$, allowing one to deduce the existence and indices of additional critical points based on partial knowledge of the invariant sets and the homology of $M$.

Morse theory has blossomed into a powerful array of topological and dynamical theories. One significant extension is the theory of Conley [14] which associates to an 'isolated' invariant set of a finite dimensional dynamical system an index – the *Conley index* – which, like the Morse index, can be used to force the existence of certain invariant sets. Instead of being a number (the dimension of the unstable manifold), the Conley index is a homotopy class of spaces (roughly speaking, the homotopy type of the unstable set). See [44] and the references therein for a sampling of applications to differential equations.

Following on the heels of Conley's index theory is the extension of Floer to infinite-dimensional gradient-like dynamics. This, in turn, has led to an explosion of results in topology and geometry. The recent flurry of activity in contact homology and symplectic field theory [18] is a descendent of these foundational ideas.

**Example 2** (*The Poincaré–Birkhoff Theorem* [5]). This theorem applies to orientation and area preserving homeomorphisms of the annulus whose boundaries are twisted in opposite directions. As with Morse theory, the forcing is in terms of a lower bound (two) on the number of fixed points. The Poincaré–Birkhoff Theorem is the first of many dynamical theorems to exploit the particular features of symplectic manifolds and maps which preserve this symplectic structure. The marriage of this type of theorem with the Morse-type forcing results is the *Arnold Conjecture*, for which Floer theory was first and most strikingly used.

There is a very powerful extension of the Poincaré–Birkhoff Theorem due to Franks [25] (Gambaudo and LeCalvez [39, App.] proved a slightly different version independently at about the same time). Franks' theorem states that if an area and orientation preserving diffeomorphism of the annulus has at least one periodic point, then it has infinitely many periodic orbits. See [26] for this and related results. Franks' Theorem is an excellent example of how a forcing theorem in dynamics often provides a sharp threshold for complicated dynamics: one simple invariant set implies the existence of infinitely many others. This principle finds its clearest exponent in the theorem of Sharkovsky.

**Example 3** (*Sharkovsky's Theorem* [48]). For continuous maps of the compact interval to the reals, this theorem gives a total ordering $\lhd$ on the periods of periodic orbits. The theorem states that if a map has an orbit of minimal period $P$ then it has periodic orbits of minimal period $Q$ for all $P \lhd Q$. That the minimal element of $\lhd$ is three has led to the popular coinage *"period three implies chaos."*

The Sharkovsky theorem is remarkable in that there are no assumptions on the systems beyond dimension and continuity. Yet, the topological datum assigned to a periodic orbit is merely the period and nothing more sophisticated. In general, the resolution with which a forcing theorem can act depends on two factors: (1) how narrowly one constrains the class of dynamical systems; and (2) what type of topological data one assigns to the invariant sets.

**1.2. Overview.** This paper motivates and describes a forcing theory developed by R. Vandervorst in collaboration with J.-B. Van den Berg and the author. In this context, the class of dynamics is, roughly speaking, scalar parabolic lattice dynamics. The topological data which drives the forcing theory is a relative Conley index for invariant sets based on the theory of *braids*.

The resulting forcing theory shares features with all three of the above examples. The index we construct – the *homotopy braid index* – is a Conley–Morse index and leads to Morse-type inequalities. The discrete version of the forcing theory is similar in spirit to LeCalvez' work on twist maps for annuli [38], [39], which itself is an elegant descendent of the Poincaré–Birkhoff Theorem. As with the Sharkovsky Theorem, we obtain a (partial) order on invariant sets. This leads to very simple conditions on invariant sets which force an infinite collection of additional invariant sets.

**1.3. Braids and braid types.** The use of braids in forcing theorems is not without precedent. There are various types of topological forcing in dimensions two and three related to braids. In the two-dimensional discrete case, one considers the isotopy class of a map relative to some periodic orbit(s): these are related to braids.

One definition of a *topological braid* on $n$ strands is a loop with basepoint in the configuration space of $n$ distinct unlabeled points in the disc $D^2$. One usually visualizes a braid as an embedding of $n$ intervals $\boldsymbol{u} = \{u^\alpha(t)\}_1^n$ into $D^2 \times [0, 1]$ such that each slice $D^2 \times \{t\}$ is a set of $n$ points and the initial and final configurations the same: $\boldsymbol{u}(0) = \boldsymbol{u}(1)$. See Figure 1 [left]. Given a braid $\boldsymbol{u}$, its *braid class* $\{\boldsymbol{u}\}$ is the equivalence class of braids *isotopic* to $\boldsymbol{u}$, that is, homotopic to $\boldsymbol{u}$ through braids, fixing the endpoints.

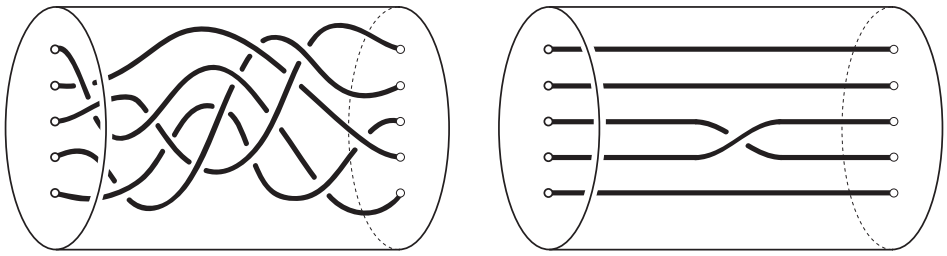


Figure 1. A braid on five strands, illustrated as a collection of embedded arcs in $D^2 \times [0, 1]$ [left]. A typical generator of the braid group has all strands 'straight' with a pair of incident strands crossing [right].

There is an obvious algebraic structure on $n$-strand braid classes by passing to the fundamental group $B_n$ of the configuration space, the group operation being concatenation of the braids in $D^2 \times [0, 1]$. The standard presentation for $B_n$ has $n - 1$ generators, where the $i^{\text{th}}$ generator consists of $n$ parallel strands (the identity braid) except that the $i^{\text{th}}$ strand crosses over the $(i + 1)^{\text{st}}$ strand as in Figure 3 [right]. See [6] for more details on the topology and algebra of braids.

There is a wonderful analogue of the Sharkovsky Theorem for forcing periodic orbits in surface homeomorphisms. In this setting, the period is not a sufficiently fine datum – one must use what Boyland [7] coined the *braid type* of a periodic orbit. Consider, for the sake of simplicity, an orientation preserving homeomorphism $f\colon D^2 \to D^2$ of the closed disc with a periodic orbit $P$ of period $n$. The braid type $\mathrm{BT}(P)$ is, roughly speaking, the isotopy class of $f$ relative to $P$. Using the relationship between braid groups and mapping class groups [6], it is possible to formally identify $\mathrm{BT}(P)$ with a conjugacy class in the braid group $B_n$ modulo its center. This is best seen by suspending the disc map to a flow on $D^2 \times S^1$. When embedded in $\mathbb{R}^3$, the periodic orbit is a braid. The choice of how many meridional twists to employ in the embedding is the genesis of modding out by the center of $B_n$.

Boyland defined the following *forcing order* on braid types: one says that $\gamma \leq \beta$ if and only if for any homeomorphism $f\colon D^2 \to D^2$ with $\gamma$ a braid type for some periodic orbit of $f$, then $\beta$ must also be a braid type for some (perhaps distinct) periodic orbit of $f$ as well. Boyland showed that this is a partial order on braid types [8], which, though weaker than the total order of the Sharkovsky theory, is nevertheless efficacious in forcing complicated dynamics.

Boyland's theory, when generalized to surfaces, entwines with the Nielsen–Thurston theory for surface homeomorphisms. This combination of braid types together with Nielsen–Thurston theory has matured to yield numerous strong results, not only in the dynamics of horseshoe and Hénon maps [11], [13], but also in problems of fluid mixing [9], [33].

**1.4. Knots and links.** In the case of flows in dimension three, embedding and isotopy data is likewise crucial. Since each periodic orbit is an embedded loop, it is a *knot*, and the set of periodic orbits forms a (perhaps infinite) *link*. The relationship between the link of periodic orbits and the dynamics of the flow is very subtle.

A forcing theory for flows in not straightforward. Roughly speaking, the counterexamples to the Seifert Conjecture constructed by K. Kuperberg [37] imply that there cannot be a forcing theorem for general smooth nonsingular 3-d flows – one can always insert a Kuperberg plug and destroy any isolated periodic orbit. At one extreme, Kuperberg's work implies that there exist smooth nonsingular flows on $S^3$ without any periodic orbits whatsoever. At the other extreme, it is possible to have a smooth, nonsingular, structurally stable flow on $S^3$ which displays all possible knots and links as periodic orbits [29]. These phenomena do not bode well for a forcing theory based on knots and links.

However, upon restriction to the correct subclass of flows, it is often possible to retain some vestige of forcing based on knot and link types. One principle which persists is that simple dynamics implicate simple types of knots. For example, in the class of nonsingular Morse–Smale flows on $S^3$, only certain knot types and link types can appear, a complete classification being given by Wada [54]. This result has a nearly dual counterpart in the class of integrable Hamiltonian dynamics on an invariant 3-sphere, as shown by Fomenko and Nguyen [24] and explained best by Casasayas et

al. [12]. Other instantiations of this principle appear in smooth, integrable fluid flows on Riemannian 3-spheres [20] and in gradient fields on $S^3$ kinematically constrained by a plane field distribution [19].

A complementary principle also holds, that complex dynamics implicate complex knot types in a flow on a 3-sphere. The best example of this type of result is the theorem of Franks and Williams [27], which states that any $C^2$ flow with positive topological entropy has a link of periodic orbits which has an infinite number of distinct knot types represented. Other results on knotting and linking for suspensions of Smale horseshoes have been proved by Holmes and Williams [35] and used to force bifurcations in Hénon maps. These results all employ the relationship between knots, links, and *closed braids* – conjugacy classes of braids in the braid group which are visualized by identifying the left and right endpoints of a braid.

**1.5. Toward higher dimensions.** Forcing theorems based on knots, links, or braids in higher dimensional dynamics seem hopeless at first: these objects fall apart in dimension higher than three. One possibility is to try to work with embedding data associated to higher-dimensional invariant sets, say spheres or tori, which can be knotted and linked in the appropriate codimension. At present, there is some initial work on braiding of 2-d invariant tori in 4-d flows [50] which may lead to a forcing theory. There is a great deal now known about the peculiar constraints of embedding spheres and tori in symplectic manifolds, but as yet without much in the way of dynamical implications.

We now turn to a braid-theoretic forcing theory for certain types of PDEs, where the stationary equation allows us to import three-dimensional embedding constraints into an infinite-dimensional dynamical system.

## 2. Braids for parabolic dynamics

Our motivation for using braids to force dynamics comes from a very simple observation about parabolic partial differential equations.

**2.1. Motivation: parabolic PDEs.** Consider the scalar parabolic PDE

$$u_t = u_{xx} + f(x, u, u_x), \tag{1}$$

where $f$ satisfies one's favorite technical assumptions to guarantee no finite-time blowups of solutions. For simplicity, we assume periodic boundary conditions ($x \in [0, 1]/0 \sim 1$). We view Equation (1) as an evolution equation on the curve $u(\cdot, t)$. As $t$ increases, the graph of $u$ evolves in the $(x, u)$ plane. Thus, the PDE induces a flow on a certain infinite-dimensional space of curves. It is a result of Fiedler and Mallet-Paret [21] that a type of Poincaré–Bendixson Theorem holds for these types of equations: the only bounded invariant sets are stationary solutions, time-periodic solutions, and connecting orbits.

We augment the types of solutions under consideration as follows. First, we allow multiple graphs to evolve by the product flow. That is, if $u^1 = u^1(t)\colon [0,1] \to \mathbb{R}$ and $u^2 = u^2(t)\colon [0,1] \to \mathbb{R}$ are solutions to Equation (1), then we consider the union $\boldsymbol{u} = (u^1, u^2)$ as a solution to the product flow. These two strands evolve together, independently, as a pair of graphs on the $(x, u)$ plane. In general, we can consider an $n$-tuple $\boldsymbol{u} = (u^k)_1^n$ of *strands* which evolve under the dynamics.

Second, we allow for strands of multiple spatial period. That is, we allow for a collection $\boldsymbol{u} = (u^k)_1^n$ of strands of the form $u^k\colon [0,1] \to \mathbb{R}$ with the endpoints equivalent as sets: $\{u^k(0)\}_1^n = \{u_k(1)\}$. Even though the endpoints do not match strandwise, the union of the endpoints of the strands do match, and thus the entire collection evolves under the PDE so as to respect the spatial periodicity. One can think of such a collection of strands as a single-strand curve on the $n$-fold cover $[0, n]/0 \sim n$ of the spatial variable $x$.

It is a well-known fact (going back to Sturm, but revived and extended considerably by Matano [41], Brunovsky and Fiedler [10], Angenent [1], and others) that there is a *comparison principle* for Equation (1). Specifically, let $u^1(t)$ and $u^2(t)$ be solutions to Equation (1). Then the number of intersections of the graphs of $u^1(t)$ and $u^2(t)$ is a weak Lyapunov function for the dynamics: it is non-increasing in $t$. Furthermore, at those particular times $t$ for which the graphs of $u^1(t)$ and $u^2(t)$ are tangent, the number of intersections decreases strictly in $t$, even in the case where the tangencies are of arbitrarily high order [1]. These facts are all at heart an application of classical maximum principle arguments which have a topological interpretation: *parabolic dynamics separates tangencies monotonically.*

This monotonicity is easily seen. Assume that $u^1$ and $u^2$ are solutions to Equation (1) which have a simple tangency where $u^1(x,t) = u^2(x,t)$. Then the evolution of the difference between $u^1$ and $u^2$ is given by

$$\frac{\partial}{\partial t}\left(u^1(x,t) - u^2(x,t)\right) = \frac{\partial^2}{\partial x^2}\left(u^1(x,t) - u^2(x,t)\right). \tag{2}$$

Since the nonlinear terms cancel, the evolution is governed purely on the basis of the curvature of the graphs.

Using this comparison principle (also known as *lap number* or *zero crossing* techniques), numerous authors have analyzed the dynamics of Equation (1) in varying degrees of generality. We note in particular the paper of Fiedler and Mallet-Paret [21], in which the comparison principle is used to show that the dynamics of Equation (1) is often Morse–Smale, and also the paper of Fiedler and Rocha [22], in which the global attractor for the dynamics is roughly classified.

**2.2. Idea: dynamics on spaces of braids.** A typical collection of strands is illustrated in Figure 2 [left], in which one notices a resemblance to the planar projection of a braid. By lifting this collection of strands in the $(x, u)$ plane to the 1-jet extension of the strands in $(x, u, u_x)$ space, we obtain a *Legendrian braid* tangent to the contact structure $\{dy - z\,dx = 0\}$. Such a braid is *closed*, due to the periodicity of

the strands. Being Legendrian, the braid is *positive* – in the standard generators for the braid group, only positive powers of generators are permitted.

There is a globalization of the comparison principle using braids. For a motivating example, consider again a pair of evolving curves $u^1(t)$ and $u^2(t)$ in the $(x, u)$ plane. If we lift these curves to the three-dimensional $(x, u, u_x)$ space, we no longer have intersecting curves, unless $t$ is such that the planar graphs of $u^1$ and $u^2$ intersect tangentially. The graphs of $u^1$ and $u^2$ in the $(x, u, u_x)$ space are instead a closed braid on two strands. What was the intersection number of their projections is now the *linking number* of the pair of strands.

We see therefore that the comparison principle takes on a linking number interpretation (a fact utilized in a discrete setting by LeCalvez [38]). After lifting solutions $u^1$ and $u^2$ to the $(x, u, u_x)$ space, the comparison principle says that the linking number is a nonincreasing function of time which decreases strictly at those times at which the curves are tangent. This two-strand example is merely motivation for adopting a braid-theoretic perspective on multiple strands, as in Figure 2.

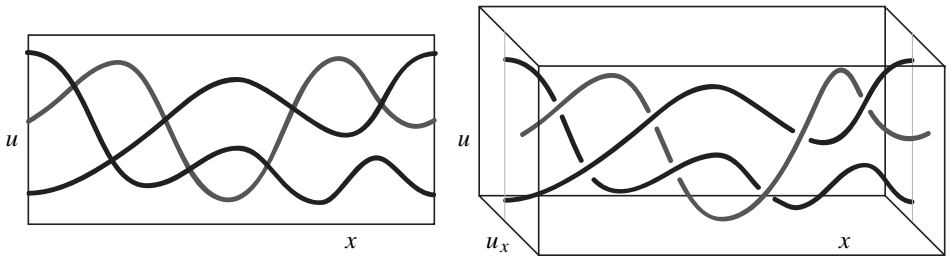



Figure 2. Curves in the $(x, u)$ plane lift to a braid in $(x, u, u_x)$.

The key observation is that the comparison principle passes from a local statement ("linking number decreases at a tangency") to a global statement ("algebraic length in the braid group decreases at a tangency"). A related globalization of the comparison principle for geodesic flows on Riemannian surfaces appears in the recent work of Angenent [2].

**2.3. Goal: forcing.** Our goal is to produce a forcing theory for the dynamics of Equation (1) and more general parabolic systems. For simplicity, we focus on forcing stationary solutions, though periodic and connecting orbits are likewise accessible. Say that one has found a *skeleton* of stationary strands $\{v^1, v^2, \ldots, v^m\}$ for a particular representative of Equation (1). How many and which types of other stationary curves are forced to be present? Since the skeleton of known fixed curves $\boldsymbol{v} = \{v^1\}_{i=1}^m$ lifts to a braid, the problem is naturally couched in braid-theoretic terms: given a braid $\boldsymbol{v}$ fixed by a particular uniform parabolic PDE, which other classes of braids $\boldsymbol{u}$ are forced to exist as stationary curves?

The spirit of our forcing theory is as follows:

1. Given a braid of stationary solutions $\boldsymbol{v}$, construct the configuration space of all $n$-strand braids $\boldsymbol{u}$ which have $\boldsymbol{v}$ as a sub-braid.

2. Use the braid-theoretic comparison principle to give a Morse-type decompose of this configuration space into dynamically isolated braid classes.

3. Define the *homotopy braid index* – a Conley index for relative braid classes which depends only on the topology of the braids, and not on the analytical details of the dynamics.

4. Prove Morse-type inequalities for forcing stationary and/or time-periodic solutions.

To execute this requires a significant generalization to spatially discretized systems, which in turn generalizes the results far beyond parabolic PDEs.

## 3. Spaces of braids for parabolic dynamics

**3.1. Braids, topological and discrete.** The motivation of §2.1 leads one to consider spaces of braids obtained from curves in the $(x, u)$ plane. Consider the space of all such $n$-strand braids $\boldsymbol{u}$ which are both closed and positive. For the sake of intuition, one should think of these topological braids as smooth braids lifted from the 1-jet extension of graphs in the plane. In reality, one completes this space to include non-smooth braids as well. These configuration spaces of braids are infinite dimensional. By projecting to finite dimensional approximations, we avoid a great deal of analytic and topological difficulties. We briefly outline the "finite dimensional" braid theory needed.

The class of *discretized braids* are best visualized as piecewise-linear braid diagrams, as in Figure 3 [left]. A discretized braid, $\boldsymbol{u}$, on $n$ strands of period $p$, is determined by *np anchor points*: $\boldsymbol{u} = \{u_i^\alpha\}_{i=0,\ldots,p}^{\alpha=1,\ldots,n}$. Superscripts $\alpha = 1, \ldots, n$ refer to strand numbers, and subscripts $i = 0, \ldots, p$ refer to spatial discretizations. One connects the anchor point $u_i^\alpha$ to $u_{i-1}^\alpha$ and $u_{i+1}^\alpha$ via straight lines. Since "height" is determined by slope, all crossings in the braid diagram are of the same sign (as in Figure 3 [left] but not in Figure 1 [left]). Since we employ periodic boundary conditions on the $x$ variable, all of the braids are closed: left and right hand endpoints of strands are abstractly identified and the endpoints are free to move. This necessitates a periodicity convention for the subscript. For a single-strand component $u^\alpha$, we have that $u_{i+p}^\alpha = u_i^\alpha$ for all $i$. For multi-strand components, one cycles between the strands according to the permutation of strands. Denote by $\mathcal{D}_p^n$ the space of all $n$-strand period $p$ discretized braids: $\mathcal{D}_p^n$ is homeomorphic to $\mathbb{R}^{np}$.

For topological braids, a *singular braid* is one for which one or more strands intersect. For braids which are lifts of graphs, the only possible intersection is that which occurs when two strands are tangent in the projection. For a discretized braid $\boldsymbol{u}$,

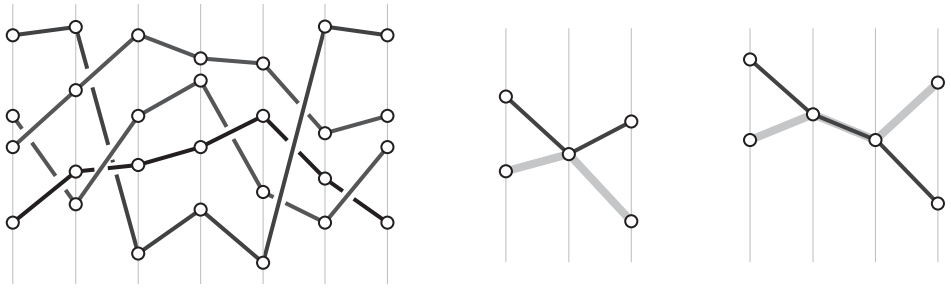


Figure 3. A discretized braid in $\mathcal{D}_6^4$ with three components (note: left and right hand sides are identified) [left]; Two types of singular discretized braids: a simple tangency, and a high-order contact [right].

the singular braids are defined to be those braids at which anchor points on two different strands coincide in a topologically non-transverse fashion with respect to immediate neighbors. Denote by $\Sigma$ the singular braids,

$$\Sigma = \{\boldsymbol{u} : u_i^\alpha = u_i^\beta \text{ for some } i \text{ and } \alpha \neq \beta, \text{ and } (u_{i-1}^\alpha - u_{i-1}^\beta)(u_{i+1}^\alpha - u_{i+1}^\beta) \geq 0\}. \tag{3}$$

The set $\Sigma$ is a discriminant that carves $\mathcal{D}_p^n$ into components: these are the *discretized braid classes*, denoted $[\boldsymbol{u}]$. Within $\Sigma$, there is a subspace of *collapsed* braids, $\Sigma^- \subset \Sigma$, consisting of those braids for which distinct components of the braid (or a single component with multiple period) collapse to yield a braid on fewer strands. More specifically,

$$\Sigma^- = \{\boldsymbol{u} \in \Sigma \, : \, u_i^\alpha = u_i^\beta \text{ for all } i \in \mathbb{Z} \text{ and some } \alpha \neq \beta\}, \tag{4}$$

under the convention of subscript periodicity mod $p$ as regulated by the braid.

**3.2. Parabolic dynamics on braids.** A parabolic PDE of the form in Equation (1) gives rise to a flow on the space of topological braids. There is likewise a broad class of flows on spaces of discretized braids which are best described as parabolic. These come from nearest-neighbor lattice dynamics.

Discretizing Equation (1) in the standard way would yield a family of nearest neighbor equations of the form $\frac{d}{dt}u_i = f_i(u_{i-1}, u_i, u_{i+1})$ in which uniform parabolicity would manifest itself in terms of the derivatives of $f_i$ with respect to the first and third variables. Instead of explicitly discretizing the PDE itself, we use the broadest possible category of nearest neighbor equations for which a comparison principle holds: these are related to the *monotone systems* of, e.g., [49], [34], [21] and others.

A *parabolic relation* of period $p$ is a sequence of maps $\mathcal{R} = \{\mathcal{R}_i : \mathbb{R}^3 \to \mathbb{R}\}$, such that $\partial_1 \mathcal{R}_i > 0$ and $\partial_3 \mathcal{R}_i \geq 0$ for every $i$. These include discretizations of uniform parabolic PDE's, as well as a variety of other discrete systems [40], [42], including

monotone twist maps [38]. The small amount of degeneracy permitted ($\partial_3 \mathcal{R}_i = 0$) does not prevent the manifestation of a comparison principle. Given a discretized braid $\boldsymbol{u} = \{u_i^\alpha\}$ and a parabolic relation $\mathcal{R}$, one evolves the braid according to the equation

$$\frac{d}{dt}(u_i^\alpha) = \mathcal{R}_i(u_{i-1}^\alpha, u_i^\alpha, u_{i+1}^\alpha). \tag{5}$$

Any parabolic relation $\mathcal{R}$ therefore induces a flow on $\mathcal{D}_p^n$. Fixed points of this flow correspond to stationary braids $\boldsymbol{u}$ satisfying $\mathcal{R}_i(u_i^\alpha) = 0$ for all $i$ and $\alpha$. It will be useful at certain points to work with parabolic relations which induce a gradient flow on $\mathcal{D}_p^n$. One calls $\mathcal{R}$ *exact* if there exist generating functions $S_i$ such that

$$\mathcal{R}_i(u_{i-1}, u_i, u_{i+1}) = \partial_2 S_{i-1}(u_{i-1}, u_i) + \partial_1 S_i(u_i, u_{i+1}), \tag{6}$$

for all $i$. In the exact case, the flow of Equation (5) is given by the gradient of $\sum_i S_i$.

All parabolic relations, exact or non-exact, possess a discrete braid-theoretic comparison principle.

**Lemma 4** (Comparison principle for braids [32]). *Let $\mathcal{R}$ be any parabolic relation and $\boldsymbol{u} \in \Sigma - \Sigma^-$ any non-collapsed singular braid. Then the flowline $\boldsymbol{u}(t)$ of $\mathcal{R}$ passing through $\boldsymbol{u} = \boldsymbol{u}(0)$ leaves a neighborhood of $\Sigma$ in forward and backward time so as to strictly decrease the algebraic length of $\boldsymbol{u}(t)$ in the braid group as $t$ increases through zero.*

Lemma 4 implies that the flow of parabolic dynamics is gradient-like on the (non-collapsed portions of) boundaries of braid classes. This suggests a Morse-theoretic approach. For example, if the flow points in to a given braid class everywhere along the boundary, then the braid class should serve as a 'sink' for the dynamics and thus be assigned a Morse index of zero. At least some invariant set would have to lie within this braid class, even if the dynamics is not gradient everywhere. For more complicated behaviors on the boundary of a braid class, Conley's version of Morse theory is the appropriate tool, with the notion of a Morse index generalizing to the Conley index, a homotopy class of spaces.

## 4. The homotopy braid index

One significant problem with this idea is the prevalence of collapsed braids, which are invariant under the flow and foil the straightforward application of Morse theory. Clearly, *any* braid class $[\boldsymbol{u}]$ borders the set of collapsed braids $\Sigma^-$ somewhere. One need simply collapse all the strands together as an extreme degeneracy.

**4.1. Relative braids.** We are therefore naturally confronted with the need for a forcing theory. Given that a particular parabolic relation possesses a stationary braid $\boldsymbol{v}$, does it force some other braid $\boldsymbol{u}$ to also be stationary with respect to the dynamics? This necessitates understanding how the strands of $\boldsymbol{u}$ braid relative to those of $\boldsymbol{v}$.

Given a discrete braid $\boldsymbol{v} \in \mathcal{D}_p^m$, consider the set of nonsingular braids

$$\left\{\boldsymbol{u} \in \mathcal{D}_p^n \ : \ \boldsymbol{u} \cup \boldsymbol{v} \in \mathcal{D}_p^{n+m} - \Sigma_p^{n+m}\right\},$$

the path components of which define the *relative braid classes* $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$. Not only are tangencies between strands of $\boldsymbol{u}$ illegal, so are tangencies with the strands of $\boldsymbol{v}$. In this setting, the braid $\boldsymbol{v}$ is called the *skeleton*. Elements within $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ are equivalent as discrete braids fixing all strands of $\boldsymbol{v}$.

In this context, it is possible to define a Conley index for certain discrete relative braid classes. To do so, it must be shown that the braid classes $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ are *isolated* in the sense that no flowlines within $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ are tangent to the boundary of this set. It follows from Lemma 4 that $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ is isolated for the flow of Equation (5) assuming that the braid class avoids the collapsed braids $\Sigma^-$. We therefore declare a braid class $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ to be *proper* if no free strands of $\boldsymbol{u}$ can "collapse" onto $\boldsymbol{v}$ or onto each other: see Figure 4. Furthermore, to ensure compactness, it is convenient to assume that the braid class $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ is *bounded* – free strands cannot wander off to $\pm\infty$.

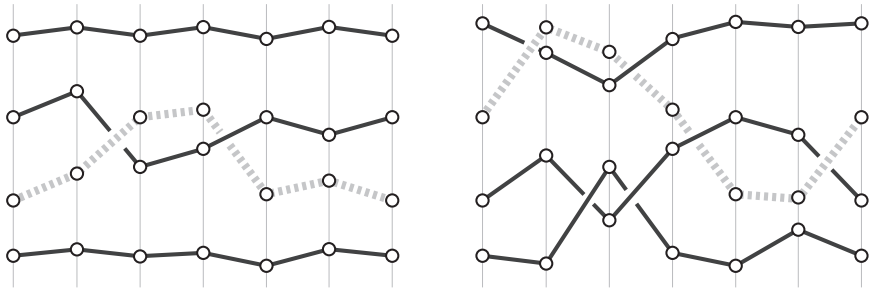


Figure 4. A bounded but improper braid class [left]. A proper, but unbounded braid class. Solid strands form the skeleton; dashed strands are free [right].

**4.2. The index: discrete version.** The *homotopy braid index* of a proper, bounded, discrete relative braid class $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ is defined as follows. Choose any parabolic relation $\mathcal{R}$ which fixes $\boldsymbol{v}$ (such an $\mathcal{R}$ exists). Define $\mathcal{E}$ to be the *exit set*: those braids on the boundary of the braid class $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ along which evolution under the flow of $\mathcal{R}$ exits the braid class. The homotopy braid index is defined to be the pointed homotopy class

$$h([\boldsymbol{u} \text{ REL } \boldsymbol{v}]) = \left(\overline{[\boldsymbol{u} \text{ REL } \boldsymbol{v}]}/\mathcal{E}, \{\mathcal{E}\}\right). \tag{7}$$

This is simply the Conley index of the closure of $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ in $\mathcal{D}_p^n$ under the flow of $\mathcal{R}$. Lemma 4, combined with the basic stability properties of the Conley index yields the following:

**Lemma 5.** *The index $h([\boldsymbol{u} \text{ REL } \boldsymbol{v}])$ is well-defined and independent of the choice of $\mathcal{R}$ (so long as it is parabolic and fixes $\boldsymbol{v}$) as well as the choice of $\boldsymbol{v}$ within its braid class $[\boldsymbol{v}]$.*

Thanks to the comparison principle for braids, the computation of the index $h$ does not require a choice of $\mathcal{R}$. One can identify the exit set $\mathcal{E}$ purely on the basis of which singular braids will decrease algebraic length under parabolic evolution. This is the basis for an algorithm to compute the homological index $H_*(h[\boldsymbol{u} \text{ REL } \boldsymbol{v}])$ numerically [17].

**Example 6.** Consider the proper period-2 braid illustrated in Figure 5 [left]. There is exactly one free strand with two anchor points (via periodicity). The anchor point in the middle, $u_1$, is free to move vertically between the fixed points on the skeleton. At the endpoints, one has a singular braid in $\Sigma$ which is on the exit set. The end anchor point, $u_0$ ($= u_2$) can freely move vertically in between the two fixed points on the skeleton. The singular boundaries are not on the exit set since pushing $u_0$ across the skeleton increases the number of crossings.

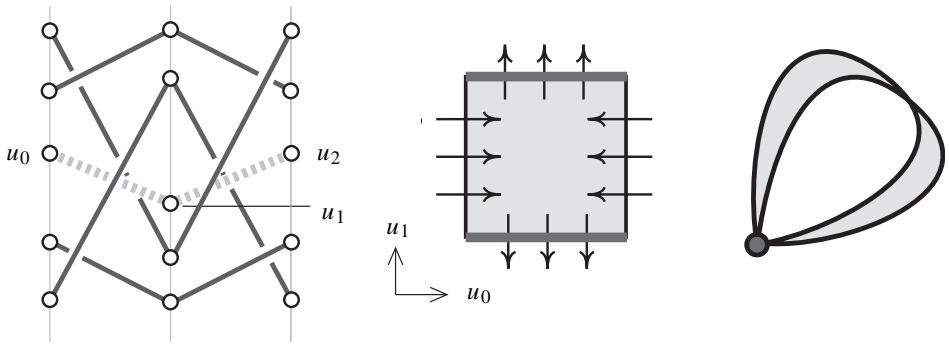



Figure 5. The braid of Example 6 [left] and the associated configuration space with parabolic flow [middle]. Collapsing out the exit set leads to a space [right] which has the homotopy type of a circle.

Since the points $u_0$ and $u_1$ can be moved independently, the braid class is the product of two intervals. The exit set consists of those points on the boundary for which $u_1$ is a boundary point. Thus, the homotopy braid index is $S^1$, as seen in Figure 5 [right].

**Example 7.** Consider the proper relative braid presented in Figure 6 [left]. Since there is one free strand of period three, the configuration space is determined by the vector of positions $(u_0, u_1, u_2)$ of the anchor points. This example differs greatly from the previous example. For instance, the point $u_0$ (as represented in the figure) may pass through the nearest strand of the skeleton above and below without changing the braid class. The points $u_1$ and $u_2$ may not pass through any strands of the skeleton

without changing the braid class unless $u_0$ has already passed through. In this case, either $u_1$ or $u_2$ (depending on whether the upper or lower strand is crossed) becomes free.

The skeleton induces a cubical partition of $\mathbb{R}^3$ by planes of singular braids. The relative braid class is the collection of cubes in $\mathbb{R}^3$ illustrated in Figure 6 [right]: it is homeomorphic to $D^2 \times S^1$. In this case, the exit set is the entire boundary and the quotient space is homotopic to the wedge-sum $S^2 \vee S^3$, the space defined by abstractly gluing a point of $S^2$ to a point of $S^3$.

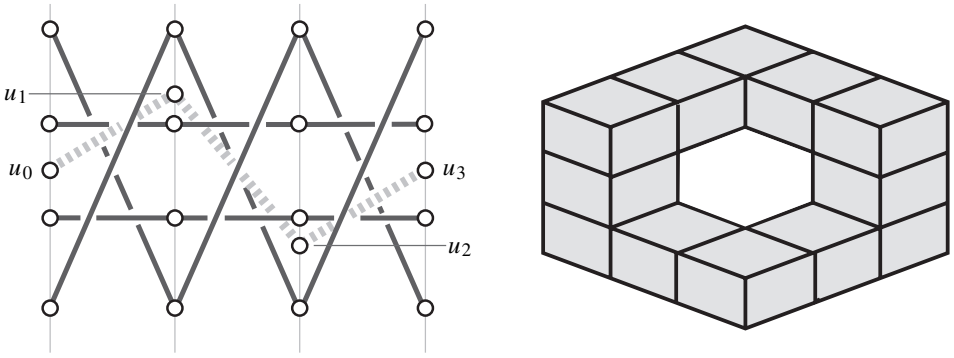



Figure 6. The braid of Example 7 and the associated relative braid class.

**Example 8.** The braid pair of Figure 7 [right] has index $h \simeq S^4 \vee S^5$ (as computed in [32, Lem. 50]); the pair on the left has trivial index, even though the linking numbers and periods of all strands are identical. This exemplifies the extra information carried by the braiding data.

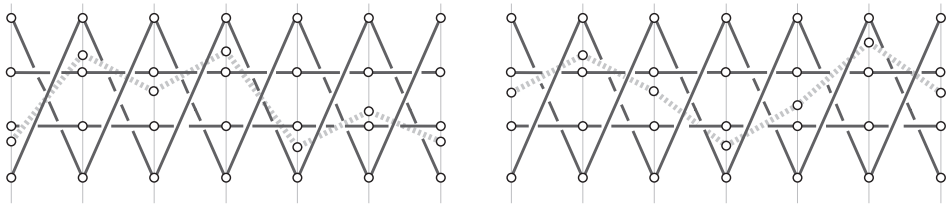


Figure 7. Discretized braid pairs with trivial [left] and nontrivial [right] homotopy index.

**4.3. The index: topological version.** As defined, the homotopy braid index $h$ is a function of discretized braid classes. For *topological* braids, one could hope that any discretization yields the same discrete index. It does, modulo two technicalities.

The first is simple. Given a topological relative braid pair $\boldsymbol{u}$ REL $\boldsymbol{v}$ and a discretization period $p$, consider the discrete braid pair whose anchor points are defined in the obvious way using $x_i = i/p$ as the spatial discretization points. Only for $p$ sufficiently large will this discrete braid pair be isotopic as a topological braid to the pair $\boldsymbol{u}$ REL $\boldsymbol{v}$. Thus, one must choose $p$ so that the correct braid class is obtained by discretization.

The second technicality is more subtle. Even if the discretized braid is topologically isotopic to the original, it is possible to "fracture" the homotopy type of the topological braid class via discretization. Consider the discrete braids of Figure 8: these braid pairs are equivalent as topological closed braids, but *not* as discrete closed braids. There is simply not enough freedom to maneuver.

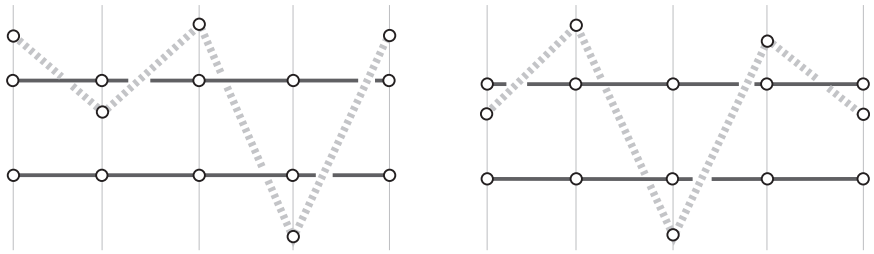


Figure 8. An example of two discretized braids which are of the same topological braid class but define disjoint discretized braid classes in $\mathcal{D}_4^1$ REL $\boldsymbol{v}$.

To overcome this difficulty, we define a modification of the homotopy braid index as follows. Given a fixed period $p$ and a discrete proper relative braid class $\beta = [\boldsymbol{u} \text{ REL } \boldsymbol{v}] \in \mathcal{D}_p^n$, let $\mathcal{S}(\beta)$ denote the set of all braid classes in $\mathcal{D}_p^n$ REL $\boldsymbol{v}$ which are isotopic as *topological* braids to a representative of $\beta$. Define the index $\boldsymbol{H}$ to be

$$\boldsymbol{H}(\beta) = \bigvee_{\beta_i \in \mathcal{S}(\beta)} h(\beta_i). \tag{8}$$

This is a wedge sum of the indices of all discrete period-$p$ representatives of the given topological braid class. The wedge sum is well-defined since each $h$ is a pointed homotopy class.

This index $\boldsymbol{H}$ is an invariant of topological braid classes. Consider the following *stabilization operator*, $\mathbb{E}\colon \mathcal{D}_p^n \to \mathcal{D}_{p+1}^n$, which appends a trivial period-1 braid to the end of a discrete braid:

$$(\mathbb{E}\boldsymbol{u})_i^\alpha = \begin{cases} u_i^\alpha, & i = 0, \ldots, p, \\ u_p^\alpha, & i = p+1. \end{cases} \tag{9}$$

The most important result about the index is the following invariance theorem:

**Theorem 9** (Stabilization [32]). *For $\boldsymbol{u}$ REL $\boldsymbol{v}$ any bounded proper discretized braid pair, the topological homotopy braid index is invariant under the extension operator:*

$$\boldsymbol{H}(\mathbb{E}\boldsymbol{u} \text{ REL } \mathbb{E}\boldsymbol{v}) = \boldsymbol{H}(\boldsymbol{u} \text{ REL } \boldsymbol{v}). \tag{10}$$

The proof of this theorem involves, surprisingly enough, a dynamical argument, utilizing a singular perturbation of a particular parabolic relation adapted to $\mathbb{E}$. This is a very convenient way to prove homotopy equivalence, given the robustness of the Conley index with respect to singular perturbations [15]. This theorem allows for a proof of topological invariance.

**Theorem 10** (Invariance [32]). *Given $\boldsymbol{u}$ REL $\boldsymbol{v} \in \mathcal{D}_p^n$ REL $\boldsymbol{v}$ and $\tilde{\boldsymbol{u}}$ REL $\tilde{\boldsymbol{v}} \in \mathcal{D}_{\tilde{p}}^n$ REL $\tilde{\boldsymbol{v}}$ which are topologically isotopic as bounded proper braid pairs, then*

$$\boldsymbol{H}(\boldsymbol{u} \text{ REL } \boldsymbol{v}) = \boldsymbol{H}(\tilde{\boldsymbol{u}} \text{ REL } \tilde{\boldsymbol{v}}). \tag{11}$$

The key ingredients in this proof are the Stabilization Theorem combined with a braid-theoretic argument that the moduli space of discretized braids converges to that of topological braids under sufficiently many applications of $\mathbb{E}$ – the length of the braid in the word metric suffices.

## 5. Forcing theorems: parabolic lattice dynamics

The dynamical consequences of the index are forcing results. A simple example: given any parabolic relation $\mathcal{R}$ which has as stationary solutions the skeleton of Figure 7 [right], then, since adding the dashed strand from that figure yields a nontrivial braid index, there must be some invariant set for $\mathcal{R}$ within this braid class. At this point, one uses Morse theory ideas: if $\mathcal{R}$ is exact, then there must be a stationary solution of the form of the grey strand. If the flow is not a gradient flow, then finer information can still detect stationary solutions.

More specifically, let $h$ be the homotopy braid index of a proper bounded discrete braid class $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$. Let $P_\tau(h)$ denote the *Poincaré polynomial* of the index – the polynomial in $\mathbb{Z}[\tau]$ whose coefficients are the Betti numbers of the homology of the index, $H_*(h;\mathbb{R})$. The following results are consequences of degenerate Morse theory (cf. [16]).

**Theorem 11** ([32]). *Given a parabolic relation $\mathcal{R}$ which fixes $\boldsymbol{v}$ and $h = h([\boldsymbol{u} \text{ REL } \boldsymbol{v}])$, the following hold:*

1. *The number of stationary braids in this braid class is bounded below by the Euler characteristic $\chi(h) = P_{-1}(h)$.*
2. *If $\mathcal{R}$ is exact, then the number of stationary braids in this braid class is bounded below by the number of nonzero monomials of $P_\tau(h)$.*

Stronger results are available if it is known that the parabolic relation is nondegenerate. By iterating the process of adding free strands and computing a nontrivial index, one can go quite far. The following forcing theorem (for exact $\mathcal{R}$) is very general, requiring only that the parabolic relation is exact (yielding a gradient flow) and *dissipative*, meaning that $\mathcal{R}_i \to -\infty$ as $|u_i| \to +\infty$.

**Theorem 12** ([32]). *Let $\mathcal{R}$ be a parabolic relation which is both exact and dissipative. If $\mathcal{R}$ fixes a discretized braid $\boldsymbol{v}$ which is not a trivial braid class, then there exist an infinite number of distinct braid classes which arise as stationary solutions of $\mathcal{R}$.*

This theorem is very much in the spirit of "period-three implies chaos." The dissipative boundary condition at infinity can be replaced with a coercive condition (infinity is attracting) or with mixtures thereof with only minor adjustments to the theorem statements [32].

## 6. Forcing theorems: second-order Lagrangians

This forcing theory gives an elegant approach to a class of fourth-order equations arising from a Lagrangian. Consider a *second order Lagrangian*, $L(u, u_x, u_{xx})$, such as is found in the Swift–Hohenberg equation:

$$L = \frac{1}{2}(u_{xx})^2 - (u_x)^2 + \frac{1-\alpha}{2}u^2 + \frac{u^4}{4}. \tag{12}$$

Assume the standard convexity assumption that $\partial^2_{u_{xx}} L \geq \delta > 0$. The Euler–Lagrange equations yield a fourth-order ODE. The objective is to find bounded functions $u : \mathbb{R} \to \mathbb{R}$ which are stationary for the action integral $J[u] = \int L(u, u_x, u_{xx})\, dx$. Due to the translation invariance $x \mapsto x + c$, the solutions of the Euler–Lagrange equation satisfy the energy constraint

$$\left(\frac{\partial L}{\partial u_x} - \frac{d}{dx}\frac{\partial L}{\partial u_{xx}}\right)u_x + \frac{\partial L}{\partial u_{xx}}u_{xx} - L(u, u_x, u_{xx}) = E = \text{ constant}, \tag{13}$$

where $E$ is the energy of a solution. To find bounded solutions for given values of $E$, we employ the variational principle $\delta_{u,T} \int_0^T (L(u, u_x, u_{xx}) + E)\, dx = 0$, which forces solutions to have energy $E$.

The Lagrangian problem can be reformulated as a two degree-of-freedom Hamiltonian system. In that context, bounded periodic solutions are closed characteristics of the corresponding energy manifold $M^3 \subset \mathbb{R}^4$. Unlike the case of first-order Lagrangian systems, the energy hypersurface is not of contact type in general [4], and is never compact. The recent stunning results in contact homology [18] are inapplicable.

**6.1. The twist condition.** The homotopy braid index provides a very effective means of forcing periodic orbits. By restricting to systems which satisfy a mild variational

hypothesis, one can employ a "broken geodesics" construction which yields a restricted form of parabolic relation.

Closed characteristics at a fixed energy level $E$ are concatenations of monotone laps between alternating minima and maxima $(u_i)_{i\in\mathbb{Z}}$, which form a periodic sequence with even period. The problem of finding closed characteristics can, in most cases, be formulated as a finite dimensional variational problem on the extrema $(u_i)$, as realized by Vandervorst, in his definition of the *twist condition*. The twist condition is a weaker version of the hypothesis that assumes that the monotone laps between extrema are unique and is valid for large classes of Lagrangians $L$, including Equation (12). The following result of [52] is the motivation and basis for the applications of the homotopy braid index to second-order Lagrangians.

**Lemma 13.** *Extremal points* $\{u_i\}$ *for bounded solutions of second order Lagrangian twist systems are solutions of an exact parabolic relation with the constraints that* (i) $(-1)^i u_i < (-1)^i u_{i+1}$*; and* (ii) *the relation blows up along any sequence satisfying* $u_i = u_{i+1}$.

**6.2. A general result.** It is necessary to retool the homotopy braid index to the setting of Lemma 13 and show that the index properties with respect to this restricted class of parabolic relations are invariant. Upon so doing, one extracts very general forcing theorems, a simple example of which is the following:

**Theorem 14** ([32]). *Let* $L(u, u_x, u_{xx})$ *be a Lagrangian which is dissipative* (*infinity is repelling*) *and twist. Then, at any regular energy level, the existence of a single periodic orbit which traces out a self-intersecting curve in the* $(u, u_x)$ *plane implies the existence of infinitely many other periodic orbits at this energy level.*

Additional results give lower bounds on the multiplicity of solutions in a given braid class based on the Poincaré polynomial and apply to singular energy levels, as well as to non-dissipative systems [32].

## 7. Forcing theorems: parabolic PDEs

The homotopy braid index, being inspired by parabolic PDEs, is efficacious in this context also, thanks to Theorem 10. By performing a spatial discretization of the dynamics of Equation (1), it is possible to reduce the dynamics of the PDE to those of a parabolic relation on a finite-dimensional space of discretized braids.

On account of the robustness of the homotopy index with respect to the dynamics, there is very little one needs to assume about the nonlinearity in Equation (1). The first, crucial, hypothesis is a growth condition on the $u_x$ term of $f$. For simplicity, let us call Equation (1) *subquadratic* if there exist constants $C > 0$ and $0 < \gamma < 2$, such that $|f(x, u, v)| \leq C(1 + |v|^\gamma)$, uniformly in both $x \in S^1$ and on compact intervals in $u$. This is necessary for regularity and control of derivatives of solution

curves, cf. [3]. This condition is sharp: one can find examples of $f$ with quadratic growth in $u_x$ for which solutions have singularities in $u_x$. Since our topological data are drawn from graphs of $u$, the bounds on $u$ imply bounds on $u_x$ and $u_{xx}$.

A second gradient hypothesis will sometimes be assumed. One says Equation (1) is *exact* if

$$u_{xx} + f(x, u, u_x) = a(x, u, u_x)\left[\frac{d}{dx}\partial_{u_x} L - \partial_u L\right], \tag{14}$$

for a strictly positive and bounded function $a = a(x, u, u_x)$ and some Lagrangian $L$ satisfying $a(x, u, u_x) \cdot \partial^2_{u_x} L(x, u, u_x) = 1$.

In this case, one has a gradient system whose stationary solutions are critical points of the action $\int L(x, u, u_x)\, dx$ over loops of integer period in $x$. This condition holds for a wide variety of systems. In general, systems with Neumann or Dirichlet boundary conditions admit a gradient-like structure which precludes the existence of nonstationary time-periodic solutions. It was shown by Zelenyak [55] that this gradient-like condition holds for many nonlinear boundary conditions which are a mixture of Dirichlet and Neumann.

**7.1. Stationary solutions.** Assume for the following theorems that $\{\boldsymbol{u} \text{ REL } \boldsymbol{v}\}$ is a topological braid class which is both bounded and proper. Assume further that $\boldsymbol{v}$ is stationary for Equation (1). We state our existence and multiplicity results in terms of the Poincaré polynomial $P_\tau(\boldsymbol{H})$ of the topological (as opposed to the discrete) braid index $\boldsymbol{H} = \boldsymbol{H}\{\boldsymbol{u} \text{ REL } \boldsymbol{v}\}$, computed via a discretization of the topological braid.

**Theorem 15** ([31]). *Let Equation* (1) *be subquadratic with* $\boldsymbol{v}$ *a stationary braid, and* $\boldsymbol{H} = \boldsymbol{H}(\{\boldsymbol{u} \text{ REL } \boldsymbol{v}\})$.

1. *There exists a stationary solution in this braid class if the Euler characteristic of the index,* $\chi(\boldsymbol{H}) = P_{-1}(\boldsymbol{H})$*, is nonvanishing.*

2. *If Equation* (1) *is furthermore exact, then there exists a stationary solution in this braid class if* $P_\tau(\boldsymbol{H}) \neq 0$.

Additional results are available concerning multiplicity of solutions, alternate boundary conditions, and non-uniformly parabolic equations: see [31]. A version of Theorem 12 on infinite numbers of braids being forced by a single nontrivial stationary braid persists in this context. The result is simplest to state if the PDE is *dissipative*; that is, $u\, f(x, u, 0) \to -\infty$ as $|u| \to +\infty$ uniformly in $x \in S^1$. This is a fairly benign restriction.

**Theorem 16** ([31]). *Let Equation* (1) *be subquadratic, exact, and dissipative. If* $\boldsymbol{v}$ *is a nontrivially braided stationary skeleton, then there are infinitely many braid classes represented as stationary solutions. Moreover, the number of single-free-strand braid classes is bounded from below by* $\lceil \iota/2 \rceil - 1$*, where* $\iota$ *is the maximal number of intersections between two strands of* $\boldsymbol{v}$.

**7.2. Examples.** The following family of spatially inhomogeneous Allen–Cahn equations was studied by Nakashima [45], [46]:

$$\varepsilon^2 u_t = \varepsilon^2 u_{xx} + g(x)u(1-u^2), \tag{15}$$

where $g\colon S^1 \to (0,1)$ is not a constant. This equation has stationary solutions $u = 0, \pm 1$ and is exact with Lagrangian

$$L = \frac{1}{2}\varepsilon^2 u_x^2 - \frac{1}{4}g(x)u^2(2-u^2).$$

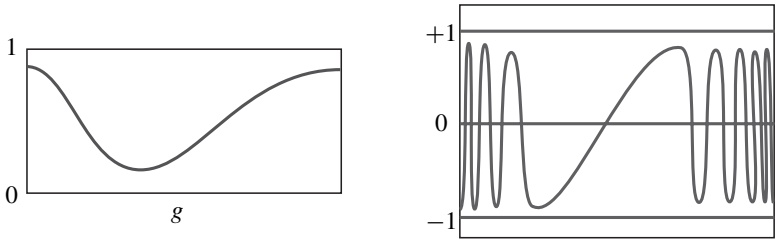



Figure 9. Given a function $g\colon S^1 \to (0,1)$ and $\varepsilon$ small, there exists a skeleton of stationary curves for Equation (15) which forms a nontrivial braid. This forces infinitely many other stationary braids.

According to [45], for any $N > 0$, there exists an $\varepsilon_N > 0$ so that for all $0 < \varepsilon < \varepsilon_N$, there exist at least two stationary solutions which intersect $u = 0$ exactly $N$ times. (The cited works impose Neumann boundary conditions: it is a simple generalization to periodic boundary conditions.) Via Theorem 16 we have that for any such $g$ and any small $\varepsilon$, this equation admits an infinite collection of stationary periodic curves; furthermore, there is a lower bound of $N$ on the number of 1-periodic solutions.

As a second explicit example, consider the equation

$$u_t = u_{xx} - \frac{5}{8}\sin 2x\, u_x + \frac{\cos x}{\cos x + \frac{3}{\sqrt{5}}}u(u^2-1), \tag{16}$$

with $x \in S^1 = \mathbb{R}/2\pi\mathbb{Z}$. This gives an exact system with Lagrangian

$$L = e^{-\frac{5}{16}\cos 2x}\left(\frac{1}{2}u_x^2 - \frac{\cos x}{\cos x + \frac{3}{\sqrt{5}}}\frac{(u^2-1)^2}{4}\right), \tag{17}$$

and weight $a(x,u,u_x) = e^{\frac{5}{16}\cos 2x}$ (cf. Equation (14)).

One checks easily that there are stationary solutions $\pm 1$ and $\pm\frac{1}{2}\big(\sqrt{5}\cos x + 1\big)$, as in Figure 10 [left]. These curves comprise a skeleton $\boldsymbol{v}$ which can be discretized to yield the skeleton of Example 6. This skeleton forces a stationary solution of the braid

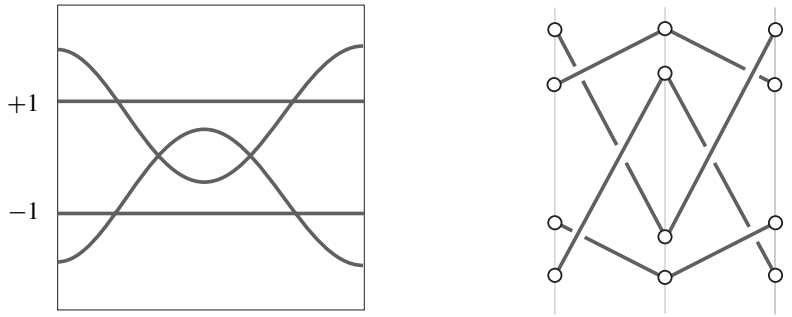



Figure 10. This collection of stationary solutions for Equation (16) [left] discretizes to the braid skeleton of Example 6.

class indicated in Figure 5 [left]: of course, this is detecting the obvious stationary solution $u = 0$. Note, however, that since $\boldsymbol{H} \simeq S^1$, this solution is unstable.

What is more interesting is the fact that one can take periodic extensions of the skeleton and add free strands in a manner which makes the relative braid spatially non-periodic. Let $\boldsymbol{v}^n$ be the $n$-fold periodic extension of $\boldsymbol{v}$ on $[0, n]/0 \sim n$ and consider a single free strand that weaves through $\boldsymbol{v}^n$ as in Figure 11. The homotopy index of such a braid is a sphere whose dimension is a function of the linking number of the free strand with the skeletal strands. The appropriate Morse inequalities imply that for each $n > 0$ there exist at least $3^n - 2$ distinct stationary solutions. This information can be used to prove that the time-$2\pi$ map of the stationary equation has positive entropy, see e.g. [47], [53].

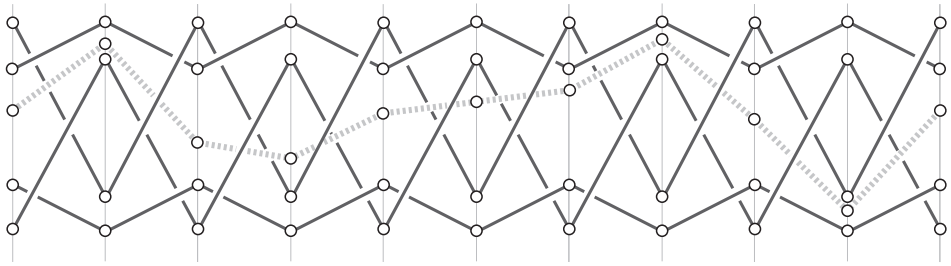


Figure 11. Taking a lift of the spatial domain allows one to weave free strands through the lifted skeleton. These project to multiply-periodic solutions downstairs. The braid pictured has index $\boldsymbol{H} \simeq S^2$.

**7.3. Time-periodic solutions.** A fundamental class of time-periodic solutions to Equation (1) are the so-called *rotating waves*. For an equation which is autonomous in $x$, one makes the rotating wave hypothesis that $u(t, x) = U(x - ct)$, where $c$ is

the unknown wave speed. Stationary solutions for the resulting equation on $U(\xi)$ yield rotating waves. In [3] it was proved that time-periodic solutions are necessarily rotating waves for an equation autonomous in $x$. However, in the non-autonomous case, the rotating wave assumption is highly restrictive.

The homotopy braid index presents a very general technique for finding time-periodic solutions without the rotating wave hypothesis.

**Theorem 17** ([31]). *Let* $\{\boldsymbol{u} \text{ REL } \boldsymbol{v}\}$ *be a bounded proper topological braid class with* $\boldsymbol{u}$ *a single-component braid,* $\boldsymbol{v}$ *an arbitrary stationary braid, and* $P_\tau(\boldsymbol{H}) \neq 0$. *If the braid class is not stationary for Equation* (1) *– the equation does not contain stationary braids in this braid class – then there exists a time-periodic solution in this braid class.*

It was shown in [3] that a singularly perturbed van der Pol equation,

$$u_t = \varepsilon u_{xx} + u(1 - \delta^2 u^2) + u_x u^2, \tag{18}$$

possesses an arbitrarily large number of rotating waves for $\varepsilon \ll 1$ sufficiently small and fixed $0 < \delta$. The homotopy braid index methods extend these results dramatically.

**Theorem 18** ([31]). *Consider the equation*

$$u_t = u_{xx} + ub(u) + u_x c(x, u, u_x), \tag{19}$$

*where c has sub-linear growth in* $u_x$ *at infinity. Moreover, b and c satisfy the following hypotheses:*

1. $b(0) > 0$, *and b has at least one positive and one negative root;*
2. $c(x, 0, 0) = 0$, *and* $c > 0$ *on* $\{uu_x \neq 0\}$.

*Then this equation possesses time-periodic solutions spanning an infinite collection of braid classes.*

All of the periodic solutions implied are dynamically unstable. In the most general case (those systems with $x$-dependence), the periodic solutions are not rigid rotating waves and thus would seem to be very difficult to detect.

## 8. What does this index mean?

The most important fact about the homotopy braid index $\boldsymbol{H}$ is that it is an invariant of topological braid pairs. Though it is not realistic to think that this is of interest in knot theory as a means of distinguishing braid pairs, the homotopy braid index nevertheless entwines both topological and dynamical content.

Thinking in terms of braid classes gives finer information than relying merely on intersection numbers. With the braid-theoretic approach, various analytic conditions

on a PDE or lattice system (dispersive, coercive, etc.) can be 'modeled' by an auxiliary braid when computing the index. Likewise, spatial boundary conditions (Neumann, Dirichlet, periodic, etc.) can be viewed as restrictions on braids (fixed, closed, etc.). Any such restrictions which yield topologically equivalent braids have the same dynamical implications with respect to forcing. One may replace complicated analytic constraints with braids.

The precise topological content to the homotopy braid index is not transparent. A few steps toward unmasking the meaning of the index are as follows.

**8.1. Duality.** One special feature of working with discretized braids in a fixed period is a natural duality made possible by the fact that the index pair used to compute the homotopy braid index can be chosen to be a manifold pair.

The *duality operator* on discretized braids of even period is the map $\mathbb{D}\colon \mathcal{D}^n_{2p} \to \mathcal{D}^n_{2p}$ given by

$$(\mathbb{D}\boldsymbol{u})^\alpha_i = (-1)^i u^\alpha_i. \tag{20}$$

Clearly $\mathbb{D}$ induces a map on relative braid diagrams by defining $\mathbb{D}(\boldsymbol{u} \text{ REL } \boldsymbol{v})$ to be $\mathbb{D}\boldsymbol{u} \text{ REL } \mathbb{D}\boldsymbol{v}$. The topological action of $\mathbb{D}$ is to insert a half-twist at each spatial segment of the braid. This has the effect of linking unlinked strands, and, since $\mathbb{D}$ is an involution, linked strands are unlinked by $\mathbb{D}$, as in Figure 12.

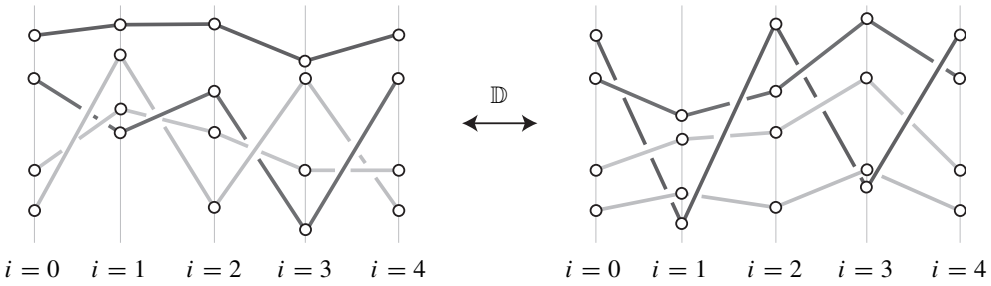



Figure 12. The topological action of $\mathbb{D}$.

For the two duality theorems to follow, we assume that all braids considered have even periods and that all of the braid classes and their duals are proper, so that the homotopy index is well-defined. In this case, the duality map $\mathbb{D}$ respects braid classes: if $[\boldsymbol{u}] = [\boldsymbol{u}']$ then $[\mathbb{D}(\boldsymbol{u})] = [\mathbb{D}(\boldsymbol{u}')]$. Bounded braid classes are taken to bounded braid classes by $\mathbb{D}$.

The effect of $\mathbb{D}$ on the index pair is to reverse the direction of the parabolic flow. This is the key to proving the following:

**Theorem 19** (Duality [32][1]). *For* $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ *having period* $2p$ *and* $n$ *free strands,*

$$H_q(\boldsymbol{H}(\mathbb{D}(\boldsymbol{u} \text{ REL } \boldsymbol{v})); \mathbb{R}) \cong H_{2np-q}(\boldsymbol{H}(\boldsymbol{u} \text{ REL } \boldsymbol{v}); \mathbb{R}). \tag{21}$$

This duality operator is very useful in computing the homology of the braid index: see the computations in [32].

**8.2. Twists.** The duality operator yields a result on the behavior of the index under appending a full twist.

**Theorem 20** (Shift [32]). *Appending a full twist to a braid shifts the homology of the index up by dimension equal to twice the number of free strands.*

We include a sketch of the proof (a more careful version of which would deal with some boundedness issues). Assume that $[\boldsymbol{u} \text{ REL } \boldsymbol{v}]$ is a braid of period $2p$ with $n$ free strands. A period two full-twist braid can be realized as the dual of the trivial braid of period two. Thus, the effect of adding a full twist to a braid can be realized by the operator $\mathbb{DEED}$. By combining Theorems 9 and 19, we obtain:

$$\begin{aligned} H_q(\boldsymbol{H}(\mathbb{DEED}[\boldsymbol{u} \text{ REL } \boldsymbol{v}])) &\cong H_{2np+2n-q}(\boldsymbol{H}(\mathbb{DEE}[\boldsymbol{u} \text{ REL } \boldsymbol{v}])) \\ &\cong H_{2np+2n-q}(\boldsymbol{H}(\mathbb{D}[\boldsymbol{u} \text{ REL } \boldsymbol{v}])) \\ &\cong H_{q-2n}(\boldsymbol{H}([\boldsymbol{u} \text{ REL } \boldsymbol{v}])). \end{aligned} \tag{22}$$

A homotopy version of Equation (22) should be achievable by following a similar procedure as in the proof of Theorem 9. We suspect one obtains an iterated suspension of the homotopy index, as opposed to a shift in homology.

## 9. Toward arbitrary braids

Given the motivation from PDEs and the comparison principle, the types of braids considered in this paper are positive braids. One naturally wonders whether an extension to arbitrary braids – those with mixed crossing types – is possible. Unfortunately, passing to discretized braids is no longer simple, as anchor points alone cannot capture crossing information for arbitrary braids.

One way to define a formal index for general braid pairs is to use Garside's Theorem [6], slightly modified. Garside's Theorem states that any braid can be placed into a unique normal form of a positive braid times a (minimal) number of negative half-twists. Clearly, one can define a modified Garside normal form that gives a unique decomposition into a positive braid and a (minimal) number of negative full twists. By applying Theorem 20, one can define a homological braid index (with negative grading permitted) by shifting the braid index of the positive normal form down by

[1]The theorem in the reference has a slight error in the statement. There, it was implicitly assumed that the braid has one free strand. The present statement is correct for arbitrary numbers of strands.

the appropriate amount. A homotopy theoretic version could be defined in terms of spectra via suspensions. This, then, yields a formal index for arbitrary (proper) braid pairs.

The real question is what dynamical meaning this generalized index entails. The passage from positive braids to arbitrary braids is akin to the passage from a Lagrangian to a Hamiltonian settings, and such an extended index appears to be a relative Floer homology for (multiply) periodic solutions to time-periodic Hamiltonian systems.

## References


[1] Angenent, S., The zero set of a solution of a parabolic equation. *J. Reine Ang. Math.* **390** (1988), 79–96.

[2] Angenent, S., Curve Shortening and the topology of closed geodesics on surfaces. *Ann. of Math.* **162** (2005), 1187–1241.

[3] Angenent, S., Fiedler, B., The dynamics of rotating waves in scalar reaction diffusion equations. *Trans. Amer. Math. Soc.* **307** (2) (1988), 545–568.

[4] Angenent, S., Van den Berg, B., Vandervorst, R., Contact and noncontact energy hypersurfaces in second order Lagrangian systems. Preprint, 2001.

[5] Birkhoff, G., Proof of Poincaré's Geometric Theorem. *Trans. Amer. Math. Soc.* **14** (1913), 14–22.

[6] Birman, J. S., *Braids, Links and Mapping Class Groups*, Ann. of Math. Stud. 82, Princeton University Press, Princeton, N.J., 1975.

[7] Boyland, P., Braid types and a topological method for proving positive entropy. Preprint, Boston University, 1984.

[8] Boyland, P., Topological methods in surface dynamics. *Topology Appl.* **58** (3) (1994), 223–298.

[9] Boyland, P., Aref, H., Stremler, M., Topological fluid mechanics of stirring. *J. Fluid Mech.* **403** (2000), 277–304.

[10] Brunovský P., Fiedler B., Connecting orbits in scalar reaction-diffusion equations. In *Dynamics Reported*, Vol. 1, Dynam. Report. Ser. Dynam. Systems Appl. 1, John Wiley & Sons, Ltd., Chichester; B. G. Teubner, Stuttgart, 1988, 57–89.

[11] de Carvalho, A., Hall, T., Pruning theory and Thurston's classification of surface homeomorphisms. *J. European Math. Soc.* **3** (4) (2001), 287–333.

[12] Casasayas, J., Martinez Alfaro, J., Nunes, A., Knots and links in integrable Hamiltonian systems. *J. Knot Theory Ramifications* **7** (2) (1998), 123–153.

[13] Collins, P., Forcing relations for homoclinic orbits of the Smale horseshoe map. *Experimental Math.* **14** (1) (2005), 75–86.

[14] Conley, C., *Isolated Invariant Sets and the Morse Index.* CBMS Reg. Conf. Ser. Math. 38, Amer. Math. Soc., Providence, R.I., 1978.

[15] Conley, C., Fife, P., Critical manifolds, travelling waves, and an example from population genetics. *J. Math. Biol.* **14** (1982), 159–176.

[16] Dancer, N., Degenerate critical points, homotopy indices and Morse inequalities. *J. Reine Angew. Math.* **350** (1984), 1–22.

[17] Day, S., Van den Berg, J., Vandervorst, R., Computing the homotopy braid index. In preparation, 2005.

[18] Eliashberg, Y., Givental, A., Hofer, H., Introduction to symplectic field theory. *Geom. Func. Anal.* Special Volume II (2000), 560–673.

[19] Etnyre, J., Ghrist, R., Gradient flows within plane fields. *Commun. Math. Helv.* **74** (1999), 507–529.

[20] Etnyre, J., Ghrist, R., Stratified integrals and unknots in inviscid flows. *Contemp. Math.*. **246** (1999), 99–112.

[21] Fiedler, B., Mallet-Paret, J., A Poincaré-Bendixson theorem for scalar reaction diffusion equations. *Arch. Rational Mech. Anal.* **107** (4) (1989), 325–345.

[22] Fiedler, B., Rocha, C., Orbit equivalence of global attractors of semilinear parabolic differential equations. *Trans. Amer. Math. Soc.* **352** (1) (2000), 257–284.

[23] Floer, A., A refinement of the Conley index and an application to the stability of hyperbolic invariant sets. *Ergodic Theory Dynam. Systems* **7** (1987), 93–103.

[24] Fomenko, A., Nguyen, T.-Z., Topological classification of integrable nondegenerate Hamiltonians on isoenergy three-dimensional spheres. In *Topological classification of integrable systems*, Adv. Soviet Math. 6, Amer. Math. Soc., Providence, RI, 1991, 267–296.

[25] Franks, J., Geodesics on $S^2$ and periodic points of annulus homeomorphisms. *Invent. Math.* **108** (1992), 403–418.

[26] Franks, J., Rotation numbers and instability sets. *Bull. Amer. Math. Soc.* **40** (2003), 263–279.

[27] Franks, J., Williams, R., Entropy and knots. *Trans. Amer. Math. Soc.* **291** (1) (1985), 241–253.

[28] Fusco, G., Oliva, W., Jacobi matrices and transversality. *Proc. Roy. Soc. Edinburgh Sect. A* **109** (1988), 231–243.

[29] R. Ghrist, Branched two-manifolds supporting all links. *Topology* **36** (2) (1997), 423–448.

[30] Ghrist, R., Holmes, P., Sullivan, M., *Knots and Links in Three-Dimensional Flows*. Lecture Notes in Math. 1654, Springer-Verlag, Berlin 1997.

[31] Ghrist, R., Vandervorst, R., Scalar parabolic PDE's and braids. Preprint, 2005.

[32] Ghrist, R., Van den Berg, J., Vandervorst, R., Morse theory on spaces of braids and Lagrangian dynamics. *Invent. Math.* **152** (2003), 369–432.

[33] Gouillart, E., Thiffeault, J.-L., Finn, M., Topological mixing with ghost rods. Preprint, 2005.

[34] M. Hirsch, Systems of differential equations which are competitive or cooperative, I: Limit sets. *SIAM J. Math. Anal.* **13** (1982), 167–179.

[35] Holmes, P., Williams, R., Knotted periodic orbits in suspensions of Smale's horseshoe: torus knots and bifurcation sequences. *Arch. Rational Mech. Anal.* **90** (2) (1985), 115–193.

[36] Kalies, W., Vandervorst, R., Closed characteristics of second order Lagrangians. Preprint, 2002.

[37] Kuperberg, K., A smooth counterexample to the Seifert conjecture. *Ann. of Math.* **140** (1994), 723–732.

[38] LeCalvez, P., Propriété dynamique des diffeomorphismes de l'anneau et du tore. *Astérique* **204** 1991.

[39] LeCalvez, P., Décomposition des difféomorphismes du tore en applications déviant la verticale. *Mém. Soc. Math. France* (*N.S.*) **79** (1999).

[40] Mallet-Paret, J., Smith, H., The Poincaré-Bendixson theorem for monotone cyclic feedback systems. *J. Dynam. Differential Equations* **2** (1990), 367–421.

[41] Matano, H., Nonincrease of the lap-number of a solution for a one-dimensional semi-linear parabolic equation. *J. Fac. Sci. Tokyo 1A* **29** (1982), 645–673.

[42] Middleton, A., Asymptotic uniqueness of the sliding state for charge-density waves. *Phys. Rev. Lett.* **68** (5) (1992), 670–673.

[43] Milnor, J., *Morse Theory*. Ann. of Math. Stud. 51, Princeton University Press, Princeton, NJ, 1963.

[44] Mischaikow, K., Conley index theory. In *Dynamical Systems* (Montecatini Terme), Lecture Notes in Math. 1609, Springer-Verlag, Berlin 1995, 119–207.

[45] Nakashima, K., Stable transition layers in a balanced bistable equation. *Differential Integral Equations* **13** (7–9) (2000), 1025–1038.

[46] Nakashima, K., Multi-layered stationary solutions for a spatially inhomogeneous Allen-Cahn equation. *J. Differential Equations* **191** (1) (2003), 234–276.

[47] Séré, E., Looking for the Bernoulli shift. *Ann. Inst. Henri Poincaré* **10** (5) (1993), 561–590.

[48] Sharkovski, A., Coexistence of cycles of a continuous map of a line to itself. *Ukranian Math. J.* **16** (1964), 61–71.

[49] Smillie, J., Competative and cooperative tridiagonal systems of differential equations. *SIAM J. Math. Anal.* **15** (1984), 531–534.

[50] Spears, B., Hutchings, M., Szeri, A., Topological bifurcations of attracting 2-tori of quasiperiodically driven nonlinear oscillators. *J. Nonlinear Sci.* **15** (6) (2005) 423–452.

[51] Sturm, C., Mémoire sur une classe d'équations à différences partielles. *J. Math. Pure Appl.* **1** (1836), 373–444.

[52] Van den Berg, J., Vandervorst, R., Fourth order conservative Twist systems: simple closed characteristics. *Trans. Amer. Math. Soc.* **354** (2002), 1383–1420.

[53] Van den Berg, J., Vandervorst, R., Wójcik, W., Chaos in orientation preserving twist maps of the plane. Preprint, 2004.

[54] Wada, M., Closed orbits of nonsingular Morse-Smale flows on $S^3$. *J. Math. Soc. Japan* **41** (3) (1989), 405–413.

[55] Zelenyak, T., Stabilization of solutions of boundary value problems for a second order parabolic equation with one space variable. *Differential Equations* **4** (1968), 17–22.

Department of Mathematics and Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, U.S.A.

E-mail: ghrist@math.uiuc.edu

# Newton interpolation polynomials, discretization method, and certain prevalent properties in dynamical systems

Anton Gorodetski, Brian Hunt*, and Vadim Kaloshin†

**Abstract.** We describe a general method of studying prevalent properties of diffeomorphisms of a compact manifold $M$, where by *prevalent* we mean true for Lebesgue almost every parameter $\varepsilon$ in a generic finite-parameter family $\{f_\varepsilon\}$ of diffeomorphisms on $M$.

Usually a dynamical property $\mathcal{P}$ can be formulated in terms of properties $\mathcal{P}_n$ of trajectories of finite length $n$. Let $\mathcal{P}$ be such a dynamical property that can be expressed in terms of only periodic trajectories. *The first idea* of the method is to *discretize* $M$ and split the set of all possible periodic trajectories of length $n$ for the entire family $\{f_\varepsilon\}$ into a *finite number* of approximating periodic pseudotrajectories. Then for each such pseudotrajectory, we estimate the measure of parameters for which it fails $\mathcal{P}_n$. This bounds the total parameter measure for which $\mathcal{P}_n$ fails by a finite sum over the periodic pseudotrajectories of length $n$. Application of Newton interpolation polynomials to estimate the measure of parameters that fail $\mathcal{P}_n$ for a given periodic pseudotrajectory of length $n$ is *the second idea*.

We outline application of these ideas to two quite different problems:

• Growth of number of periodic points for prevalent diffeomorphisms (Kaloshin–Hunt).

• Palis' conjecture on finititude of number of "localized" sinks for prevalent surface diffeomorphisms (Gorodetski–Kaloshin).

**Mathematics Subject Classification (2000).** 37C05, 37C50, 37D25, 37C29.

**Keywords.** Discretization method, Newton interpolation polynomials, prevalence, pseudotrajectory, growth of number of periodic points, Newhouse phenomenon.

## 1. Introduction

A classical problem in dynamics, geometry, and topology is the description of generic behavior. Given a set of objects what are the properties of a generic element of the set? This question applies to diffeomorphisms, Riemannian metrics, linear operators, and vector fields, just to give several examples. The traditional approach is based on the category theorem of Baire. A countable intersection of open, dense sets is called a *residual,* or *topologically generic*, set. The Baire category theorem says that topologically generic sets of a complete metric space (or, more generally, Baire space) are dense. The book of Oxtoby [O] provides a rich variety of topologically generic mathematical objects. However, in many different areas of mathematics examples

*Supported by NSF grant DMS0104087.

†Supported by an Alfred Sloan Research Fellowship, American Institute of Mathematics Fellowship, and NSF Grant No. DMS-0300229

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

of "wild behavior" of topologically generic objects have been detected (see [HSY], [Ka2], [OY], [Si] and references there). In this paper we are concerned with generic properties in dynamics, particularly those that are not generic topologically but are generic in a measure-theoretic sense.

In the 1960s two main theories in dynamical systems were developed, one of which was designed for conservative systems and called *KAM* for *Kolmogorov–Arnold–Moser* and the other was constructed for general dynamical systems (nonconservative, dissipative) and called *hyperbolic*.

Kolmogorov [Ko], in his plenary talk of ICM 1954, pointed out that a different notion of genericity may be appropriate: "In order to obtain negative results concerning insignificant or exceptional character of some phenomenon we shall apply the following, somewhat haphazard, technique: if in a class $K$ of functions $f(x)$ one can introduce a finite number of functionals

$$F_1(f),\ F_2(f), \ldots,\ F_r(f),$$

which in some sense can naturally be considered as taking "arbitrary" values in general

$$F_1(f) = C_1,\ F_2(f) = C_2, \ldots, F_r(f) = C_r$$

from some domain of the $r$-dimensional space of points $C = (C_1, \ldots, C_r)$, then any phenomenon that takes place only if $C$ belongs to a set of zero $r$-dimensional measure will be regarded exceptional and subject to "neglect"."

A somewhat similar way to define a measure-theoretic genericity, often called *prevalence*, is the following: *We call a property $\mathcal{P}$ prevalent if for a generic*[1] *finite-parameter family $\{f_\varepsilon\}_{\varepsilon\in B}$ for Lebesgue almost every parameter $\varepsilon$ the corresponding $f_\varepsilon$ satisfies $\mathcal{P}$. If complement of a property is prevalent such a property is called shy*. We shall discuss prevalence further in Section 9.

There are many examples when topological genericity and measure-theoretic genericity do not coincide. We just mention a few of them (see [HSY], [Ka2], [OY] for many more).

• *Diophantine numbers* form a set of full measure on the line $\mathbb{R}$, but are topologically negligible (that is the complement of the set is topologically generic).

• For a topologically generic, even open dense, set of circle maps preserving orientation there is a finite number of attracting and repelling periodic orbits. All other orbits accumulate to these orbits both forward or backward in time. However, as the famous example of Arnold, called *Arnold tongues*, shows in the family $f_{\alpha,\varepsilon}\colon \theta \mapsto \theta + \alpha + \varepsilon \sin\theta$ that the smaller $\varepsilon$ is, the smaller is the measure of $\alpha$ values such that $f_{\alpha,\varepsilon}$ has this property. Moreover, the main result of KAM theory says that for consevative systems close to integrable most, in a measure-theoretic sense, motions are *quasiperiodic*.

• In general dynamical systems a dream of the 1960s was to prove that a generic dynamical system is structurally stable. However, this dream evaporated by the end of

[1] We give a rigorous definition in Section 9.

that decade. One of the beautiful counterexamples is due to Newhouse [N1], [N2]. He shows that there is an open set in the space of diffemorphisms of a compact manifold such that a generic diffeomorphism in this open set has *infinitely many coexisting sinks* (attracting periodic orbits). Below we show in some weak sense this phenomenon is shy (see Section 7). This phenomenon is closely related to Palis' program [Pa] which is discussed next.

Let $\mathrm{Diff}^r(M)$ be the space of $C^r$ diffeomorphisms of a smooth compact manifold $M$ with the uniform $C^r$-topology, where $\dim M \geq 2$, and let $f \in \mathrm{Diff}^r(M)$. The main focus of the present paper is *the space of general* (*nonconservative*) *diffeomorphisms* $\mathrm{Diff}^r(M)$. The authors believe that the method presented here also applies to conservative systems.

While examples such as Newhouse's show that on open subsets of $\mathrm{Diff}^r(M)$, "wild" phenomena that are not structurally stable can be topologically generic, a measure-theoretic point of view may be more appropriate to describe the dynamical behavior that would typically be observed by a scientist. In the influential paper J. Palis [Pa] proposed a new global view of generic dynamics based on measure theory. He stated the following conjectures on finititude of attractors and their metric stability:

(I) *Denseness of finititude of attractors – there is* $C^r$ $(r \geq 1)$ *dense set* $D$ *of diffeomorphisms in* $\mathrm{Diff}^r(M)$ *such that each element of* $D$ *has finitely many attractors, the union of whose basins of attraction has full measure;*

(II) *Existence of physical* (*SRB*) *mesure – each attractor of an element of* $D$ *supports a physical measure that represents the limiting distribution for Lebesgue almost every initial condition in its basin;*

(III) *Metric stability of basins of attraction – for each element in* $D$ *and each of its attractors, for almost all small* $C^r$ *perturbations in generic* $k$*-parameter families of diffeomorphisms in* $\mathrm{Diff}^r(M)$, $k \in \mathbb{N}$, *there are finitely many attractors whose union of basins is nearly equal in the sense of Lebesgue measure to the basin of the initial attractor; such perturbed attractors support a physical measure.*

Such results have been established for certain examples of dynamical systems. Lyubich [Ly] for the quadratic family of 1-dimensional maps and Avila–Lyubich–de Melo [ALM] for a generic family of analytic unimodal 1-dimensional maps showed that for almost all parameters the attractors are either periodic sinks or carry an absolutely continuous invariant measure. For the 1-dimensional Schrödinger cocycles Avila–Krikorian [AK] showed that for all analytic or $C^\infty$ potentials and almost all rotation numbers the corresponding cocycle is either non-uniformly hyperbolic or reducible.

In this paper we discuss *two* important topologically negligible dynamical properties that are in fact prevalent. One property is (stretched) exponential growth of the number of periodic points and the other is finiteness of number of coexisting "localized" sinks for surface diffeomorphisms.

We hope that the method, outlined in this article, brings a better understanding of prevalent properties of $\mathrm{Diff}^r(M)$ in the direction of Palis' conjectures and other important dynamical properties.

## 2. Elementary events and a sample result

Here we expose ideas in a general setting. Consider a family of diffeomorphisms $\{f_\varepsilon\}_{\varepsilon\in B} \subset \mathrm{Diff}^r(M)$ of a compact manifold with a probability measure $\mu$ supported on the set of parameters $B$. To avoid distracting details we postpone specification of $\mu$ and $B$.

Let us fix a certain property $\mathcal{P}$ of periodic points of period $n$. In both cases that we will consider, $\mathcal{P}$ is some form of quantitative hyperbolicity. We split the problem into two parts.

• Estimate the measure of the set

$$\mu(B_n) \le \mu_n,\ B_n = \{\varepsilon \in B :\ f_\varepsilon \text{ has a periodic orbit that does not satisfy } \mathcal{P}\} \subset B.$$

• Derive some dynamically interesting properties from this estimate.

The second part essentially depends on the problem. As for the first part, application of the discretization method and Newton interpolation polynomials give a uniform approach to get a required estimate. First, we discuss the problem of growth of the number of periodic points (see Theorem 2.1 below).

For $\gamma > 0$ we say that $x = f^n(x)$ is $(n,\gamma)$-*hyperbolic* if all eigenvalues of the linearization $df^n(x)$ are at least $\gamma$-away from the unit circle[2]. For $\gamma > 0$ this is a weak analog of Kupka–Smale property. Fix some $c > 0$ and a decaying to zero sequence of positive numbers $c\Gamma = \{c\gamma_n\}_{n\in\mathbb{Z}_+}$.

We say that the map $f_\varepsilon$ satisfies *the inductive hypothesis of order n with constants* $c\Gamma$, denoted $f_\varepsilon \in IH(n, c\Gamma)$, if for all $k \le n$ all periodic orbits of period $k$ are $(k, c\gamma_k)$-hyperbolic. Consider a sequence of "bad" sets in the parameter space

$$B_n(c\Gamma) = \{\varepsilon \in B :\ f_\varepsilon \in IH(n-1, c\Gamma),\ \text{but } f_\varepsilon \notin IH(n, c\Gamma)\}. \tag{1}$$

In other words, $B_n(c\Gamma)$ is the set of "bad" parameter values $\varepsilon \in B$ for which all periodic points with period strictly less than $n$ are sufficiently hyperbolic, but there is a periodic point of period $n$ that is not $(n, c\gamma_n)$-hyperbolic.

Our goal is to find an upper bound

$$\mu\{B_n(c\Gamma)\} \le \mu_n(c\Gamma) \tag{2}$$

for the measure of the set of "bad" parameter values. Then the sum over $n$ of (2) gives an upper bound $\mu\big\{\bigcup_n B_n(c\Gamma)\big\} \le \sum_{n\ge1} \mu_n(c\Gamma)$ on the set of all parameters $\varepsilon$ for which $f_\varepsilon$ has a periodic point of some period $n$ that is not $(n, c\gamma_n)$-hyperbolic. If the sum converges and $\sum_{n\ge1} \mu_n(c\Gamma) = \mu(c) \to 0$ as $c \to 0$, then for $\mu$-almost every $\varepsilon$ there is $c > 0$ such that for every $n$ every periodic point of period $n$ is $(n, c\gamma_n)$-hyperbolic.

This statement (almost) implies that all periodic points of period $n$ are at least $\approx c\gamma_n$-apart and, therefore, the number of periodic points is bounded by $\approx (c\gamma_n)^{-\dim M}$

[2] In [KH1] we use a stronger property of hyperbolicity of periodic points (see Section 2 of that paper).

(see [KH1], Proposition 1.1.6). Thus, the key to prove a statement that a certain property is prevalent, i.e. holds for almost every parameter value, is *an estimate of the probability* (2) *of a "bad" event*. One could replace the property of hyperbolicity of periodic points by another property and still the key is to get an estimate of the probability to fail a certain dynamical property.

Our goal is to outline the proof of the following result:

**Theorem 2.1** ([KH1], [Ka3], [Ka4]). *For a prevalent set of diffeomorphisms* $f \in \mathrm{Diff}^r(M)$*, with* $1 < r < \infty$*, and for all* $\delta > 0$ *there exists* $C = C(\delta)$ *such that*

$$P_n(f) := \#\{\text{ isolated } x \in M : \ f^n(x) = x\} \le \exp(Cn^{1+\delta}).$$

Density of diffeomorphisms with this property is the classical result of Artin–Mazur [AM] (see also [Ka2] for a simple proof). In [Ka1], using [GST], it is shown that diffeomorphisms having an arbitrary ahead given growth along a subsequence are topologically generic.

In Section 7 we briefly describe application of the method of the paper to Newhouse phenomenon from [GK].

## 3. Strategy to estimate probability of a "bad" event: discretization method

The goal of this section is to outline how one can get estimate (2). Usually we do not know where is a "bad" trajectory, which fails $\mathcal{P}$, and what are the dynamics in its neighborhood. So our analysis will be *implicit*. More exactly, we shall consider all possible trajectories in the family $\{f_\varepsilon\}_{\varepsilon\in B}$ and the worst case scenario for each of them.

In order to fail the inductive hypothesis of order $n$ with constants $c\Gamma$, a diffeomorphism $f_\varepsilon$ should have a periodic, but not $(n, c\gamma_n)$-hyperbolic point $x = f_\varepsilon^n(x)$. There is a continuum of possible $n$-tuples $\{x_k\}_{0\le k\le n}$ such that for some $\varepsilon \in B$ we have $f(x_k) = x_{k+1 \pmod n}$ and $x_0$ is not $(n, c\gamma_n)$-hyperbolic. Instead of looking at the continuum of $n$-tuples, we discretize this space and consider only those $n$-tuples $\{x_k\}_{0\le k\le n}$ that lie on a particular grid, denoted $I_{\tilde\gamma_n}$, and replace trajectories by $\tilde\gamma_n$-pseudotrajectories. If we choose the grid spacing $\tilde\gamma_n$ small enough, then every (almost) periodic point of period $n$ that is not sufficiently hyperbolic will have a corresponding $\tilde\gamma_n$-pseudotrajectory of length-$n$ on the grid that also has small hyperbolicity. In this way we reduce the problem of bounding the measure of a set of "bad" parameters corresponding to a particular length-$n$ $\tilde\gamma_n$-pseudotrajectory on the chosen grid.

Thus, the basic requirement for the grid size $\tilde\gamma_n$ is that every real periodic trajectory $\{x_k = f_\varepsilon^k(x_0)\}_{0\le k\le n}$ of length $n$ can be approximated by a $\tilde\gamma_n$-pseudotrajectory $\{\tilde x_k\}_{0\le k\le n}$ so that if $x_0$ is periodic but not $(n, c\gamma_n)$-hyperbolic, then the $n$-tuple $\{\tilde x_k\}_{0\le k\le n}$ is not $(n, c\gamma_n/2)$-hyperbolic (see [KH1], sect. 3.2 and [GK], sect. 8 for various definitions).

We call an $n$-tuple $\{x_k\}_{k=0}^{n-1} \subset I_{\tilde\gamma_n}^n$ *a $\tilde\gamma_n$-pseudotrajectory associated to some $\varepsilon$ (or to the map $f_\varepsilon$)* if for each $k = 0, \dots, n-1$ we have $dist(f_\varepsilon(x_{k-1}),\ x_k) \le \tilde\gamma_n$ and we call it *a $\tilde\gamma_n$-pseudotrajectory associated to $B$ (or the family $\{f_\varepsilon\}_{\varepsilon\in B}$)* if it is associated to some $\varepsilon \in B$.

The naive idea of estimate (2) consists of two steps:
*Step* 1. Estimate the number of different $\tilde\gamma_n$-pseudotrajectories $\#_n(\tilde\gamma_n)$ associated to $B$;
*Step* 2. For an $n$-tuple $\{x_k\}_{0\le k\le n-1} \subset I_{\tilde\gamma_n}^n$ estimate the measure

$$\begin{aligned}&\mu\{\varepsilon \in B :\ \{x_k\}_{0\le k\le n-1} \text{ is a } \tilde\gamma_n\text{-pseudotrajectory associated to } \varepsilon\\ &\text{which is } \tilde\gamma_n\text{-periodic but not } (n, c\gamma_n/2)\text{-hyperbolic}\} \le \mu_n(c\gamma_n, \tilde\gamma_n).\end{aligned} \tag{3}$$

Then the product of two numbers $\#_n(\tilde\gamma_n)$ and $\mu_n(c\gamma_n, \tilde\gamma_n)$ that are obtained in Steps 1 and 2 gives the required estimate. In fact, this simpleminded scheme requires modifications discussed at the end of the next section (see (10–13)).

We start with the second step. For simplicity we shall discuss 1-dimensional maps (see [KH1], sect. 3). In [KH1], sect. 4.2 we discuss difficulties arising to extend this method to multidimensional maps. See also [GK], sect. 10 (resp. [Ka4], sect. 7–8), where 2-dimensional (resp. $N$-dimensional) case is considered. To treat the multidimensional case one use very similar ideas, however, technical difficulties arising due to multidimensionality are fairly involved. Now we show how to estimate probability (3) within a particular polynomial family and then show how to do Step 1 and incorporate the method into the global framework.

## 4. Newton interpolation polynomials and an estimate of probability of a $\tilde\gamma_n$-periodic but not $(n, c\gamma_n/2)$-hyperbolic $\tilde\gamma_n$-pseudotrajectory of length $n$

Let $M$ be an interval $[-1, 1]$ and $I_{\tilde\gamma_n} \subset [-1, 1]$ be a $\tilde\gamma_n$-grid. Fix an $n$-tuple of points $\{x_k\}_{k=0}^{n-1} \subset I_{\tilde\gamma_n}$. Consider the following $2n$-parameter family of maps:

$$f_u(x) = f(x) + \sum_{k=0}^{2n-1} u_k \prod_{j=0}^{k-1} (x - x_{j \pmod n}).$$

This family is nothing but the Newton interpolation polynomials associated to the $n$-tuple $\{x_k\}_{k=0}^{n-1}$. Denote $\phi_u(x) = \sum_{k=0}^{2n-1} u_k \prod_{j=0}^{k-1}(x - x_{j \pmod n})$. Notice that

$$\begin{aligned}\phi_u(x_0) &= u_0,\\ \phi_u(x_1) &= u_0 + u_1(x_1 - x_0),\\ \phi_u(x_2) &= u_0 + u_1(x_2 - x_0) + u_2(x_2 - x_0)(x_2 - x_1),\\ &\vdots\end{aligned}$$

$$
\begin{aligned}
&\vdots\\
\phi_u(x_{n-1}) &= u_0 + u_1(x_{n-1} - x_0) + \dots\\
&\quad + u_{n-1}(x_{n-1} - x_0)\dots(x_{n-1} - x_{n-2}),\\
\phi'_u(x_0) &= \frac{\partial}{\partial x}\Big(\sum_{k=0}^{2n-1} u_k \prod_{j=0}^{k}(x - x_{j(\mathrm{mod}\ n)})\Big)\Big|_{x=x_0},\\
&\vdots\\
\phi'_u(x_{n-1}) &= \frac{\partial}{\partial x}\Big(\sum_{k=0}^{2n-1} u_k \prod_{j=0}^{k}(x - x_{j(\mathrm{mod}\ n)})\Big)\Big|_{x=x_{n-1}}.
\end{aligned}
\tag{4}
$$

These formulas are very useful for dynamics. For a given map $f$ and an initial point $x_0$, the image $f_u(x_0) = f(x_0) + \phi_u(x_0)$ of $x_0$ depends only on $u_0$. Furthermore the image can be set to any desired point by choosing $u_0$ appropriately – we say then that it depends only and nontrivially on $u_0$. If $x_0$, $x_1$, and $u_0$ are fixed, the image $f_u(x_1)$ of $x_1$ depends only on $u_1$, and as long as $x_0 \neq x_1$ it depends nontrivially on $u_1$. More generally for $0 \leq k \leq n-1$, if distinct points $\{x_j\}_{j=0}^{k}$ and coefficients $\{u_j\}_{j=0}^{k-1}$ are fixed, then the image $f_u(x_k)$ of $x_k$ depends only and nontrivially on $u_k$.

Suppose now that an $n$-tuple of pairwise distinct points $\{x_j\}_{j=0}^{n-1}$ and Newton coefficients $\{u_j\}_{j=0}^{n-1}$ are fixed. Then derivative $f'_u(x_0)$ at $x_0$ depends only and nontrivially on $u_n$. Likewise for $0 \leq k \leq n-1$, if distinct points $\{x_j\}_{j=0}^{n-1}$ and Newton coefficients $\{u_j\}_{j=0}^{n+k-1}$ are fixed, then the derivative $f'_u(x_k)$ at $x_k$ depends only and nontrivially on $u_{n+k}$.

As Figure 1 illustrates, these considerations show that for any map $f$ and any desired trajectory of distinct points with any given derivatives along it, one can choose Newton coefficients $\{u_k\}_{k=0}^{2n-1}$ and explicitly construct a map $f_u = f + \phi_u$ with such a trajectory. While the parametrization depends on the $n$-tuple, the family is equivalent by a change of parameter coordinates (see Section 5) to the family $\{f_\varepsilon\}_\varepsilon$ of perturbations by degree $2n-1$ polynomials, given by (14).

Using these properties of Newton interpolation polynomials we can easily estimate probability (3). Let us split this compound dynamic event into simple ones and use the above properties:

$$
\begin{aligned}
&1. && |f_\varepsilon(x_0) - x_1| \leq \tilde{\gamma}_n;\\
&2. && |f_\varepsilon(x_1) - x_2| \leq \tilde{\gamma}_n;\\
&&& \vdots\\
&n. && |f_\varepsilon(x_{n-1}) - x_0| \leq \tilde{\gamma}_n;\\
&n+1. && \Big|\prod_{j=0}^{n-1} |f'_\varepsilon(x_j)| - 1\Big| \leq c\gamma_n/2.
\end{aligned}
\tag{5}
$$

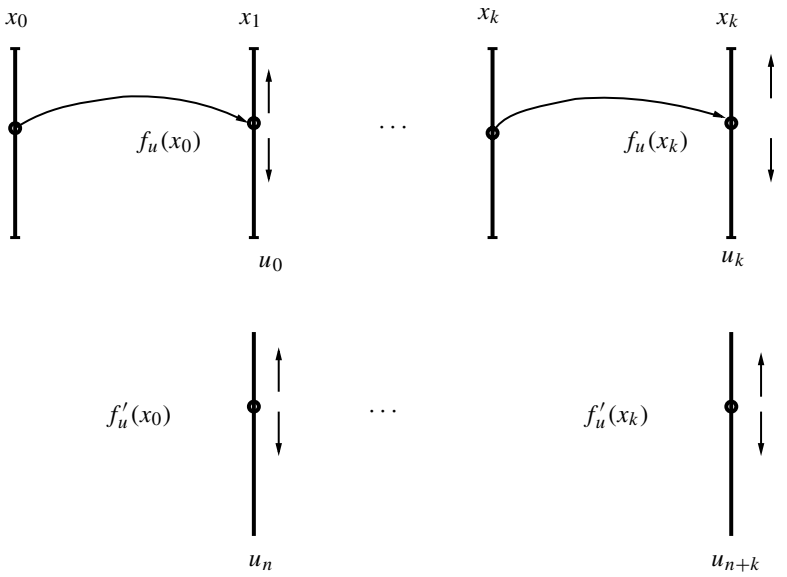



Figure 1. Newton coefficients and their action.

First, we find probabilities of these events with respect to $u$-parameters (see [KH1], sect. 3.3 for more details). It turns out that the map relating $\varepsilon$-parameters and $u$-parameters is one-to-one, linear, and volume-preserving (see Section 5).

Notice that in (4) and Figure 1, the image $f_u(x_0)$ of $x_0$ is independent of $u_k$ for all $k > 0$. Therefore, the position of $f_u(x_0)$ *depends only on* $u_0$. For the 1-dimensional Lebesgue measure of the $u_0$'s we have

$$\text{Leb}\,\{u_0 :\ |f_u(x_0) - x_1| = |f(x_0) + u_0 - x_1| \le \tilde{\gamma}_n\} \le 2\tilde{\gamma}_n.$$

Fix $u_0$. Similarly, the position of $f_u(x_1)$ *depends only on* $u_1$ (see (4) and Figure 1). Thus, we have

$$\text{Leb}\big\{u_1 :\ |f_u(x_1) - x_2| = |f(x_1) + u_0 + u_1(x_1 - x_0) - x_2| \le \tilde{\gamma}_n\} \le \frac{2\tilde{\gamma}_n}{|x_1 - x_0|}.$$

Inductively for $k = 2, \dots, n-1$, fix $u_0, \dots, u_{k-1}$. Then the position of $f_u(x_k)$ *depends only on* $u_k$. Moreover, for $k = 2, \dots, n-2$ we have

$$\begin{aligned}\text{Leb}\Big\{u_k :\ |f_u(x_k) - x_{k+1}| = \Big|f(x_k) + \sum_{m=0}^{k} u_m \prod_{j=0}^{m-1}(x_k - x_j) - x_0\Big| \le \tilde{\gamma}_n\Big\} \\ \le \frac{2\tilde{\gamma}_n}{\prod_{j=0}^{k-1}|x_k - x_j|},\end{aligned}$$

and for $k = n-1$ we have

$$\mathrm{Leb}\big\{u_{n-1} :\ |f_u(x_{n-1}) - x_0| \le \tilde{\gamma}_n\big\} \le \frac{2\tilde{\gamma}_n}{\prod_{j=0}^{n-2} |x_{n-1} - x_j|}. \tag{6}$$

In particular, the parameter $u_{n-1}$ *is responsible* for $(n, \gamma_n)$-periodicity of the $n$-tuple $\{x_k\}_{0\le k\le n}$. This formula estimates the *"measure of periodicity"*.

Choose $u_0, \dots, u_{n-1}$ so that the $n$-tuple $\{x_k\}_{k=0}^{n-1}$ is a $(n, \tilde{\gamma}_n)$-periodic $\tilde{\gamma}_n$-pseudo-trajectory. Notice that parameters $u_n, u_{n+1}, \dots, u_{2n-1}$ do not change the $\tilde{\gamma}_n$-pseudo-trajectory $\{x_k\}_{k=0}^{n-1}$. Fix now parameters $u_0, \dots, u_{2n-2}$ and vary only $u_{2n-1}$. Then for any $C^1$-smooth map $g\colon I \to I$, consider the 1-parameter family

$$g_{u_{2n-1}}(x) = g(x) + (x - x_{n-1}) \prod_{j=0}^{n-2} (x - x_j)^2.$$

Since the corresponding monomial $(x-x_{n-1}) \prod_{j=0}^{n-2}(x-x_j)^2$ has zeroes of the second order at all points $x_k$, except the last one $x_{n-1}$, we have

$$\prod_{j=0}^{n-1} (g_{u_{2n-1}})'(x_j) = \Big(g'(x_{n-1}) + u_{2n-1} \prod_{j=0}^{n-2} |x_{n-1} - x_j|^2\Big) \prod_{j=0}^{n-2} g'(x_j). \tag{7}$$

To get the final estimate, we use the fact that we are interested only in maps from the family $\{f_u\}_u$. Suppose $|f_u'(x_{n-1})|$ is uniformly bounded by some $M_1$. For condition $(n+1)$ of (5) to hold, $\big|\prod_{j=0}^{n-1} f_u'(x_j)\big|$ must lie in $[1 - c\gamma_n/2, 1 + c\gamma_n/2]$. If this occurs for any $u_{2n-1}$, then $\big|\prod_{j=0}^{n-2} f_u'(x_j)\big| \ge (1 - c\gamma_n/2)/M_1$ for all $u_{2n-1}$, because this product does not depend on $u_{2n-1}$. Using (7) and the fact that $1 - c\gamma_n/2 \ge 1/2$, we get

$$\mathrm{Leb}\left\{u_{2n-1} :\ \Big|\prod_{j=0}^{n-1} |f_u'(x_j)| - 1\Big| \le \frac{c\gamma_n}{2}\right\} \le M_1\, \frac{2c\gamma_n}{\prod_{j=0}^{n-2} |x_{n-1} - x_j|^2}. \tag{8}$$

This formula estimates the *"measure of hyperbolicity"*.

We can combine all these estimates and get

$$\begin{aligned} &\mathrm{Leb}^{n+1}\{(u_0, \dots, u_{n-1}, u_{2n-1}) : f_u \text{ satisfies conditions (5) and } \|f_u\|_{C^1} \le M_1\} \\ &\qquad \le \frac{2M_1 c\gamma_n}{\prod_{j=0}^{n-2} |x_{n-1} - x_j|^2} \prod_{m=1}^{n-1} \frac{2\tilde{\gamma}_n}{\prod_{j=0}^{m-1} |x_m - x_j|}. \end{aligned} \tag{9}$$

This completes Step 2, but leaves many open questions which we shall discuss while treating Step 1. The estimate of Step 1 then breaks down as follows:

$$\#_n(\tilde{\gamma}_n) \approx \boxed{\begin{array}{c}\#\text{ of initial}\\ \text{points in } I_{\tilde{\gamma}_n}\end{array}} \times \boxed{\begin{array}{c}\#\text{ of } \tilde{\gamma}_n\text{-pseudotrajectories}\\ \text{per initial point}\end{array}} \tag{10}$$

And up to an exponential function of $n$, the estimate of Step 2 breaks down like:

$$\mu_n(c\gamma_n, \tilde{\gamma}_n) \approx \frac{\boxed{\text{Measure of periodicity (6)}} \times \boxed{\text{Measure of hyperbolicity (8)}}}{\boxed{\#\text{ of } \tilde{\gamma}_n\text{-pseudotrajectories per initial point}}}$$

(Roughly speaking, the terms in the numerator represent respectively the measure of parameters for which a given initial point will be $(n, \tilde{\gamma}_n)$-periodic and the measure of parameters for which a given $n$-tuple is $(n, c\gamma_n)$-hyperbolic; they correspond to estimates (6) and (8) in the next section.) Thus after cancellation, the estimate of the measure of "bad" set $B_n(c\Gamma)$ associated to almost periodic, not sufficiently hyperbolic trajectories becomes:

$$\boxed{\text{Measure of bad parameters}} \leq \boxed{\#\text{ of initial points of } I_{\tilde{\gamma}_n}} \times \boxed{\text{Measure of periodicity (6)}} \times \boxed{\text{Measure of hyperbolicity (8)}} \quad (11)$$

*Consider only pseudotrajectories having* $\prod_{j=0}^{n-2} |x_{n-1} - x_j| \geq (c\gamma_n)^{1/4}$ *and suppose* $\tilde{\gamma}_n = M_1^{-n} c\gamma_n$. Then up to exponential function of $n$ the first term on the right hand side of (11) is of order $(c\gamma_n)^{-1}$. The second term has an upper bound of order $(c\gamma_n)^{3/4}$, and the third term is at most of order $(c\gamma_n)^{1/2}$, so that the product on the right-hand side of (11) is of order at most $(c\gamma_n)^{1/4}$ (up to an exponential function in $n$). If $c\gamma_n$ is exponentially small with a large exponent in $n$, then $\mu_n(c\gamma_n, \tilde{\gamma}_n)$ is at most exponentially small. This discussion motivates the following

**Definition 4.1.** A trajectory $x_0, \ldots, x_{n-1}$ of length $n$ of a diffeomorphism $f \in \mathrm{Diff}^r(B^N)$, where $x_k = f^k(x_0)$, is called $(n, \gamma)$-*simple* if

$$\prod_{k=0}^{n-2} |x_{n-1} - x_k| \geq \gamma^{1/4}.$$

A point $x_0$ is called $(n, \gamma)$-simple if its trajectory $\{x_k = f^k(x_0)\}_{k=0}^{n-1}$ of length $n$ is $(n, \gamma)$-simple. Otherwise a point (resp. a trajectory) is called non-$(n, \gamma)$-simple.

If a trajectory is simple, then perturbation of this trajectory by Newton interpolation polynomials is effective. The product of distances is a quantitative characteristic of recurrent properties of a trajectory. If it is small enough, then there are close returns of it to $x_0$ before time $n$.

Even though most of properties of periodic orbits do not depend on a starting point, it turns out that for the above product, even asymptotically, it *does* matter where to choose the starting point. A good example to look at is periodic trajectories in a neighborhood of a planar homoclinic tangency (see [KH1], sect. 2.4 for more). It motivates the following

**Definition 4.2.** A point $x$ is called *essentially* $(n, \gamma)$-simple if for some nonnegative $j < n$, the point $f^j(x)$ is $(n, \gamma)$-simple. Otherwise a point is called essentially non-$(n, \gamma)$-simple.

In (11) we consider only $(n, c\gamma_n)$-simple pseudotrajectories. To study nonsimple pseudotrajectories we look for their simple almost periodic parts. More exactly, *for each non-$(n, c\gamma_n)$-simple pseudotrajectory we find such a close return, say $x_k$, that $\{x_j\}_{j=0}^{n-1}$ is almost equal to $n/k$ copies of $\{x_j\}_{j=0}^{k-1}$ and $\{x_j\}_{j=0}^{k-1}$ is $(k, c\gamma_k)$-simple*. Due to closeness, sufficient hyperbolicity of $\{x_j\}_{j=0}^{k-1}$ implies sufficient hyperbolicity $\{x_j\}_{j=0}^{n-1}$. Then investigation of the measure of nonhyperbolicity of nonsimple pseudotrajectory reduces to the measure of nonhyperbolicity of its simple almost periodic parts. Thus to obtain $\mu_n(c\gamma_n, \tilde{\gamma}_n)$ from (3) we arrive at the following scheme:

| Measure of bad parameters associated to periodic nonhyperbolic orbits | = | | (12) |
|---|---|---|---|
| Measure of bad parameters associated to simple periodic nonhyperbolic orbits (I) | + | Measure of bad parameters associated to nonsimple periodic nonhyperbolic orbits (II) | |

| Measure of bad parameters associated to nonsimple periodic nonhyperbolic orbits | ≤ | | (13) |
|---|---|---|---|
| Partition of nonsimple periodic orbits into simple almost periodic parts (II.A) | & | Measure of bad parameters associated to short non-simple almost periodic nonhyperbolic orbits (II.B) | |

As a matter of fact (13) requires additional comments, since the left hand side is a number, while the right hand side is not. To estimate the number from the left hand side we do two step procedure described in the right hand side. First, we do a certain partition (II.A) and then estimate a different number (II.B), which turn out to be an upper bound for the left hand side.

This diagram summarizes the problems we face in the proof.

• Part (I): how to estimate the measure of parameter values (11) associated with simple periodic nonhyperbolic orbits;

• Part (II.A): how to partition a nonsimple periodic orbit into almost periodic parts so that hyperbolicity of an almost periodic part implies hyperbolicity of the whole orbit;

The part (II.B) (how to estimate the measure associated with (11) simple periodic nonhyperbolic shorter orbits) can be treated in the same way as part (I), even though the actual details are usually quite involved (see [KH1], sect. 3.5–3.6).

## 5. How to collect all simple (almost) periodic pseudotrajectories: the Distortion and Collection Lemmas

In this section for the model family we show how one can justify heuristic estimates (10 – 11). The model family is the family of perturbations of a $C^2$ map $f\colon I \to I$, $I = [-1, 1]$ such that $f(I)$ strictly belongs to $I$

$$f_\varepsilon(x) = f(x) + \sum_{k=0}^{2n-1} \varepsilon_k x^k, \quad \varepsilon = (\varepsilon_0, \dots, \varepsilon_{2n-1}). \tag{14}$$

This is a $2n$-parameter family. Assume that parameters belong to a brick, called *the brick of standard thickness* with width $\tau$ (see [KH1], sect. 3.1 in the 1-dimensional case, [KH1], sect. 4.3, [Ka4], sect. 8.3 in the $N$-dimensional case, [GK], sect. 2.3 and 11.2 for modified definitions in the 2-dimensional case applicable to the problem of finiteness of localized sinks)

$$HB^{\mathrm{st}}_{<2n}(\tau) = \left\{ \{\varepsilon_k\}_{k=0}^{2n-1} : \text{ for all } 0 \le k < 2n, \ |\varepsilon_k| < \frac{\tau}{k!} \right\}.$$

For small enough $\tau$ the map $f_\varepsilon\colon I \to I$ is well defined for all $\varepsilon \in HB^{\mathrm{st}}_{<2n}(\tau)$. Since we are interested in the measure 0 or 1 events, one could chop a brick of another shape into smaller bricks of standard thickness and use the same proof. Suppose $\sup_{\varepsilon \in HB^{\mathrm{st}}_{<2n}(\tau)} \|f_\varepsilon\|_{C^1} < M_1$ for some $M_1$.

Define the Lebesgue product probability measure, denoted by $\mu^{\mathrm{st}}_{<2n,\tau}$, on the Hilbert brick of parameters $HB^{\mathrm{st}}_{<2n}(\tau)$ by normalizing the 1-dimensional Lebesgue measure along each component to the 1-dimensional Lebesgue probability measure

$$\mu^{\mathrm{st}}_{m,\tau} = \left(\frac{m!}{2\tau}\right) \mathrm{Leb}_1, \quad \mu^{\mathrm{st}}_{<k,\tau} = \bigtimes_{m=0}^{k-1} \mu^{\mathrm{st}}_{m,\tau}.$$

By definition of $\mu^{\mathrm{st}}_{<2n,\tau}$ we have that $\varepsilon_0, \dots, \varepsilon_{2n-1}$ are independent uniformly distributed random variables.

How to get from this family to a "generic finite-parameter family" is a tedious two step procedure based on Fubini theorem. The first step, from finite-parameter polynomial families to families of analytic perturbations, is discussed in [KH1], sect. 2.3, see also [GK], sect. 3.2. The second step, from analytic perturbations to prevalent finite-parameter families, is discussed in [KH1] Appendix C.

Consider an ordered $n$-tuple of points $\boldsymbol{X}_n = \{x_k\}_{k=0}^{n-1} \in I^n$. One can define an linear map $\mathcal{L}^1_{\boldsymbol{X}_n}\colon \mathbb{R}^{2n}_\varepsilon \to \mathbb{R}^{2n}_u$ given implicitly by the following formulas

$$\sum_{k=0}^{2n-1} \varepsilon_k x^k = \sum_{k=0}^{2n-1} u_k \prod_{j=0}^{k-1} (x - x_{j\,(\mathrm{mod}\ n)}), \tag{15}$$

where $\mathcal{L}^1_{\boldsymbol{X}_n}(\varepsilon_0, \dots, \varepsilon_{2n-1}) = (u_0, \dots, u_{2n-1})$. In [KH1], sect. 2.2 we give an explicit definition of this map using so-called *divided differences*, and call it *Newton*

*map*. It provides relation between $\varepsilon$-coordinates and $u$-coordinates. It turns out that $\mathcal{L}^1_{\boldsymbol{X}_n}$ is *volume-preserving and* $\mu^{\mathrm{st}}_{<2n,\tau}$*-preserving* ([KH1] Lemma .2.2.2). *Therefore, estimate* (9) *in u-space and* $\varepsilon$*-space are the same*.

We now estimate the distortion of the Newton map $\mathcal{L}^1_{\boldsymbol{X}_n}$ as a map from the standard basis $\{\varepsilon_k\}_{k=0}^{2n-1}$ in the space of polynomials of degree $< 2n$ to the Newton basis $\{u_k\}_{k=0}^{2n-1}$. It helps to have in mind the following picture characterizing the distortion of the Newton map.

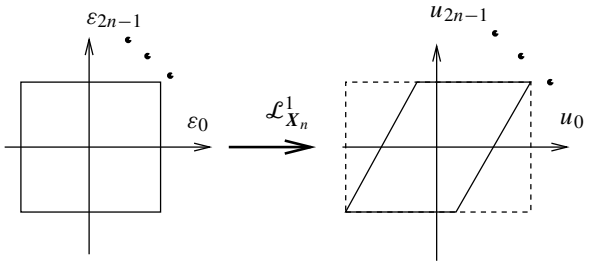



Figure 2. Distortion by the Newton map.

**The Distortion Lemma** ([KH1], sect 3.4). *Let* $\boldsymbol{X}_n = \{x_j\}_{j=0}^{n-1} \in I^n$ *be an ordered n-tuple of points in the interval* $I = [-1, 1]$ *and* $\mathcal{L}^1_{\boldsymbol{X}_n} : \mathbb{R}^{2n}_{\varepsilon} \to \mathbb{R}^{2n}_{u}$ *be the Newton map, defined implicitly by* (15). *Then the image of the brick of standard thickness* $HB^{\mathrm{st}}_{<2n}(\tau)$ *with width* $\tau > 0$ *is contained in the brick of standard thickness* $HB^{\mathrm{st}}_{<2n}(3\tau)$ *with width* $3\tau$:

$$\mathcal{L}^1_{\boldsymbol{X}_n}(HB^{\mathrm{st}}_{<2n}(\tau)) \subset HB^{\mathrm{st}}_{<2n}(3\tau) \subset \mathbb{R}^{2n}_{u}.$$

*In other words, independently of the choice of an n-tuple* $\{x_j\}_{j=0}^{n-1} \in I^n$ *for any* $0 \le m < 2n$, *the coefficient* $u_m$ *has at most the range of values* $|u_m| \le \frac{3\tau}{m!}$ *in the image* $\mathcal{L}^1_{\boldsymbol{X}_n}(HB^{\mathrm{st}}_{<2n}(\tau))$.

The proof is simple, provided the Newton map is explicitly defined (see [KH1], sect. 2.2).

In the $N$-dimensional case the statement of the necessary Distortion Lemma is somewhat involved. Even to define the $N$-dimesional Newton map one has to incorporate many multiindices (see [KH1], sect. 4.2–4.3, [Ka4], sect. 8.2–8.3) For the statement and the proof of a modified Distortion Lemma applicable to the problem of finiteness of localized sinks see [GK], sect. 11.4.

For a given $n$-tuple $\boldsymbol{X}_n = \{x_j\}_{j=0}^{n-1} \in I^n$, the parallelepiped

$$\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau) := \mathcal{L}^1_{\boldsymbol{X}_n}(HB^{\mathrm{st}}_{<2n}(\tau)) \subset \mathbb{R}^{2n}_{u}$$

is the set of parameters $(u_0, \dots, u_{2n-1})$ that correspond to parameters $(\varepsilon_0, \dots, \varepsilon_{2n-1}) \in HB^{\mathrm{st}}_{<2n}(\tau)$. In other words, these are the Newton parameters *allowed by the family* (14) for the $n$-tuple $\boldsymbol{X}_n$. Since $\mathcal{L}^1_{\boldsymbol{X}_n}$ is volume-preserving it follows that $\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)$ has the same volume as $HB^{\mathrm{st}}_{<2n}(\tau)$, but the Distortion Lemma tells us in addition that the projection of $\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)$ onto any coordinate axis is at most a factor of 3 longer than the projection of $HB^{\mathrm{st}}_{<2n}(\tau)$.

Let $\boldsymbol{X}_m = \{x_j\}_{j=0}^{m-1}$ be the $m$-tuple of first $m$ points of the $n$-tuple $\boldsymbol{X}_n$. We now consider which Newton parameters are allowed by the family (14) when $\boldsymbol{X}_m$ is fixed but $x_m, \dots, x_{n-1}$ are arbitrary. Since we will only be using the definitions below for discretized $n$-tuples $\boldsymbol{X}_n \in I^n_{\tilde{\gamma}_n}$, we consider only the (finite number of) possibilities $x_m, \dots, x_{n-1} \subset I_{\tilde{\gamma}_n}$. Let

$$\pi^{u,\boldsymbol{X}_n}_{<2n,\le m} : \mathbb{R}^{2n}_u \to \mathbb{R}^m_u \quad \text{and} \quad \pi^{u,\boldsymbol{X}_n}_{<2n,m} : \mathbb{R}^{2n}_u \to \mathbb{R}_{u_m}$$

be the natural projections onto the space $\mathbb{R}^m_u$ of polynomials of degree $m$ and the space $\mathbb{R}_{u_m}$ of homogeneous polynomials of degree $m$ respectively. Denote the unions over all $x_m, \dots, x_{n-1} \in I_{\tilde{\gamma}_n}$ of the images of $\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)$ under the respective projections $\pi^{u,\boldsymbol{X}_n}_{<2n,\le m}$ and $\pi^{u,\boldsymbol{X}_n}_{<2n,m}$ by

$$\mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau) = \bigcup_{x_m,\dots,x_{n-1}\in I_{\tilde{\gamma}_n}} \pi^{u,\boldsymbol{X}_n}_{<2n,\le m}(\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)) \subset \mathbb{R}^m_u,$$
$$\mathcal{P}^{\mathrm{st}}_{<2n,m,\boldsymbol{X}_m}(\tau) = \bigcup_{x_m,\dots,x_{n-1}\in I_{\tilde{\gamma}_n}} \pi^{u,\boldsymbol{X}_n}_{<2n,m}(\mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)) \subset \mathbb{R}_{u_m}.$$

For each $m < n$, the set $\mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau)$ is a polyhedron and $\mathcal{P}^{\mathrm{st}}_{<2n,m,\boldsymbol{X}_m}(\tau)$ is a segment of length at most $6\tau/m!$ by the Distortion Lemma. Both depend only on the $m$-tuple $\boldsymbol{X}_m$ and width $\tau$. The set $\mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau)$ consists of all Newton parameters $\{u_j\}_{j=0}^m \in \mathbb{R}^m_u$ that are allowed by the family (14) for the $m$-tuple $\boldsymbol{X}_m$.

For each $m < n$, we introduce the family of diffeomorphisms

$$f_{u(m),\boldsymbol{X}_m}(x) = f(x) + \sum_{s=0}^{m} u_s \prod_{j=0}^{s-1}(x - x_j), \tag{16}$$

where $u(m) = (u_0, \dots, u_m) \in \mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau)$. For each possible continuation $\boldsymbol{X}_n$ of $\boldsymbol{X}_m$, the family $f_{u(m),\boldsymbol{X}_m}$ includes the subfamily of $f_{u,\boldsymbol{X}_n}$ (with $u \in \mathcal{P}^{\mathrm{st}}_{<2n,\boldsymbol{X}_n}(\tau)$) corresponding to $u_{m+1} = u_{m+2} = \cdots = u_{2n-1} = 0$. However, the action of $f_{u,\boldsymbol{X}_n}$ on $x_0, \dots, x_m$ doesn't depend on $u_{m+1}, \dots, u_{2n-1}$, so for these points the family $f_{u(m),\boldsymbol{X}_m}$ is representative of the entire family $f_{u,\boldsymbol{X}_n}$. This motivates the definition

$$T^{1,\tilde{\gamma}_n}_{<2n,\le m,\tau}(f; x_0, \dots, x_{m-1}, x_m, x_{m+1})$$
$$= \big\{u(m) \in \mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau) \subset \mathbb{R}^m_u :$$
$$|f_{u(m),\boldsymbol{X}_m}(x_{j-1}) - x_j| \le \tilde{\gamma}_n \text{ for } j = 1, \dots, m+1\big\}.$$

The set $T^{1,\tilde{\gamma}_n}_{<2n,\le m,\tau}(f; x_0, \dots, x_{m-1}, x_m, x_{m+1})$ represents the set of Newton parameters $u(m) = \big(u_0, \dots, u_m\big)$ allowed by the family (14) for which $x_0, \dots, x_{m+1}$ is a $\tilde{\gamma}_n$-pseudotrajectory of $f_{u(m),\boldsymbol{X}_m}$ (and hence of $f_{u,\boldsymbol{X}_n}$ for all valid extensions $u$ and $\boldsymbol{X}_n$ of $u(m)$ and $\boldsymbol{X}_m$).

In the following lemma, we collect all possible $\tilde{\gamma}_n$-pseudotrajectories and estimates of "bad" measure corresponding to those $\tilde{\gamma}_n$-pseudotrajectories. The idea of the proof of this lemma is the following. Let $m$ be some number $0 \le m < n$. Suppose an $(m+1)$-tuple $x_0, \dots, x_m \subset I_{\tilde{\gamma}_n}$ is fixed and we are interested in the number of possible continuations $x_{m+1} \in I_{\tilde{\gamma}_n}$ so that $x_0, \dots, x_{m+1}$ is associated to the family (14). Consider the family (16), where $u_0, \dots, u_{m-1}$ are fixed. By Distortion Lemma we have $|u_m| \le \dfrac{3\tau}{m!}$. Rewrite this family, applied to $x_m$, as

$$f_{u(m),\boldsymbol{X}_m}(x_m) = f(x_m) + \sum_{s=0}^{m-1} u_s \prod_{j=0}^{s-1}(x_m - x_j) + u_m \prod_{j=0}^{m-1}(x_m - x_j).$$

Since all $u_k$'s except $u_m$ are fixed, the range of $x_{m+1}$ associated to the family (14) is bounded by $3\tau \prod_{j=0}^{m-1}(x_m - x_j)/m!$.

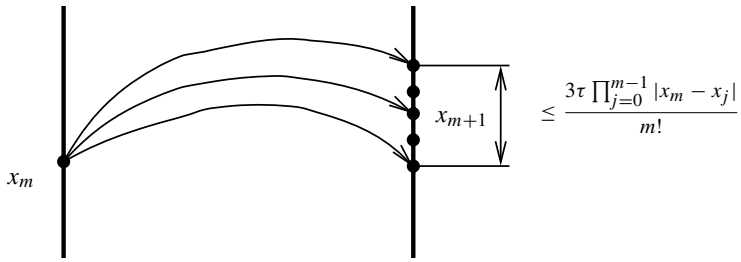



Figure 3. Collection of pseudotrajectories.

**The Collection Lemma.** *With the notations above, for all $x_0 \in I_{\tilde{\gamma}_n}$ the measure of the "bad" parameters satisfies*

$$\begin{aligned}&\mu^{\mathrm{st}}_{<2n,\tau}\big\{\varepsilon : \text{ there is a } \tilde{\gamma}_n\text{-periodic } \tilde{\gamma}_n\text{-pseudotrajectory from } I^n_{\tilde{\gamma}_n} \text{ starting at } x_0,\\ &\text{which is not } (n, M_1^{3n} c\gamma_n)\text{-hyperbolic}\big\} \le 6^{2n} M_1^{4n+1} \frac{(n-1)!}{\tau} \frac{(2n-1)!}{\tau} c^{1/4} \tilde{\gamma}_n \gamma_n^{1/4}.\end{aligned} \tag{17}$$

For the modified Collection Lemma for the $N$-dimensional case see [Ka4], sect. 9.3, and for the problem of finiteness of localized sinks in the 2-dimensional case see [GK], sect. 11.6, respectively.

**Corollary 5.1.** *With the notations above the measure of the "bad" parameters satisfies*

$$\begin{aligned}&\mu^{st}_{<2n,\tau}\big\{\varepsilon : \text{ there is a } \tilde{\gamma}_n\text{-periodic } \tilde{\gamma}_n\text{-pseudotrajectory from } I^n_{\tilde{\gamma}_n},\\ &\text{which is not } (n, M_1^{3n} c\gamma_n)\text{-hyperbolic}\big\} \le 2 \cdot 6^{2n} M_1^{4n+1} \frac{(n-1)!}{\tau} \frac{(2n-1)!}{\tau} c^{1/4} \gamma_n^{1/4}.\end{aligned}$$

Since there are $2/\tilde{\gamma}_n$-grid points of $I_{\tilde{\gamma}_n} \subset [-1, 1]$, this corollary follows directly from the Collection Lemma. Suppose that $\tilde{\gamma}_n$-discretization is fine enough to be able to approximate "real" trajectories by $\tilde{\gamma}_n$-pseudotrajectories well enough (see [KH1], Proposition 3.1.2, its proof, and (3.17) in the 1-dimensional case, [Ka4], sect. 9 in the $N$-dimensional case, and [GK], sect. 8 for the problem of finiteness of localized sinks). Then up to the error term $6^{2n} M_1^{4n+1}$ this proves (11).

*Proof of the Collection Lemma.* We prove by backward induction on $m$ that for $x_0, \dots, x_m \subset I_{\tilde{\gamma}_n}$,

$$\begin{aligned}
&\mu^{\mathrm{st}}_{<2n,\tau}\big\{\text{there is a } \tilde{\gamma}_n\text{-periodic } \tilde{\gamma}_n\text{-pseudotrajectory from } I^n_{\tilde{\gamma}_n} \text{ starting} \\
&\qquad\qquad \text{with } x_0, \dots, x_m \text{ which is not } (n, M_1^{3n} c\gamma_n)\text{-hyperbolic}\big\} \qquad (18)\\
&\le 6^{2n-m} M_1^{4n+1} \, \frac{(n-1)!}{\tau} \, \frac{(2n-1)!}{\tau} \, \mu^{\mathrm{st}}_{<m,\tau}\big\{T^{1,\tilde{\gamma}_n}_{<2n,\le m-1,\tau}(f; x_0, \dots, x_m)\big\} \, c^{1/4} \, \tilde{\gamma}_n \, \gamma_n^{1/4},
\end{aligned}$$

resulting when $m = 0$ in (17).

Consider the case $m = n - 1$. Fix an $(n, c\gamma_n/2)$-simple $n$-tuple $\boldsymbol{X}_n = \{x_j\}_{j=0}^{n-1} \in I^n_{\tilde{\gamma}_n}$. Using formulas (6) and (8), we get

$$\begin{aligned}
&\mu^{\mathrm{st}}_{n-1,\tau}\{u_{n-1} : \big|f_{u,\boldsymbol{X}_n}(x_{n-1}) - x_0\big| \le \tilde{\gamma}_n\} \\
&\qquad \le \frac{(n-1)!}{\tau} \frac{\tilde{\gamma}_n}{\prod_{m=0}^{n-2} |x_{n-1} - x_m|} \le \frac{2^{1/4}(n-1)!}{\tau} \, c^{-1/4} \, \tilde{\gamma}_n \, \gamma_n^{-1/4}
\end{aligned}$$

and

$$\begin{aligned}
&\mu^{\mathrm{st}}_{2n-1,\tau}\Big\{u_{2n-1} : \Big|\prod_{j=0}^{n-1} |(f_{u,\boldsymbol{X}_n})'(x_j)| - 1\Big| \le M_1^{3n} c\gamma_n\Big\} \\
&\qquad \le M_1 \frac{(2n-1)!}{\tau} \, \frac{4M_1^{3n} c\gamma_n}{\prod_{m=0}^{n-2} |x_{n-1} - x_m|^2} \le \frac{2^{5/2} M_1^{3n+1} (2n-1)!}{\tau} \, c^{1/2} \, \gamma_n^{1/2}.
\end{aligned}$$

The Fubini Theorem, preservation of generalized volume by the Newton map (see [KH1], Lemma 2.2.2), and the definition of the product measure $\mu^{\mathrm{st}}_{<2n,\tau}$ imply that

$$\begin{aligned}
&\mu^{\mathrm{st}}_{<2n,\tau}\big\{\text{there is a } \tilde{\gamma}_n\text{-periodic } \tilde{\gamma}_n\text{-pseudotrajectory from } I^n_{\tilde{\gamma}_n} \text{ starting} \\
&\qquad\qquad \text{with } x_0, \dots, x_{n-1} \text{ which is not } (n, M_1^{3n} c\gamma_n)\text{-hyperbolic}\big\} \\
&\le \mu^{\mathrm{st}}_{<n-1,\tau}\Big\{T^{1,\tilde{\gamma}_n}_{<2n,\le n-2,\tau}(f; x_0, \dots, x_{n-1})\Big\} \times \mu^{\mathrm{st}}_{n-1,\tau}\big\{u_{n-1} : \big|f_{u,\boldsymbol{X}_n}(x_{n-1}) - x_0\big| \le \tilde{\gamma}_n\big\} \\
&\qquad \times \prod_{s=n}^{2n-2} \mu^{\mathrm{st}}_{s,\tau}\{\mathcal{P}^{\mathrm{st}}_{<2n,s,\boldsymbol{X}_n}(\tau)\} \times \mu^{\mathrm{st}}_{2n-1,\tau}\Big\{u_{2n-1} : \Big|\prod_{j=0}^{n-1} |(f_{u,\boldsymbol{X}_n})'(x_j)| - 1\Big| \le M_1^{3n} c\gamma_n\Big\} \\
&\le 2^{11/4} 3^{n-1} M_1^{4n+1} \frac{(n-1)!}{\tau} \frac{(2n-1)!}{\tau} \, \mu^{\mathrm{st}}_{<n-1,\tau}\big\{T^{1,\tilde{\gamma}_n}_{<2n,\le n-2,\tau}(f; x_0, \dots, x_{n-1})\big\} \, c^{1/4} \, \tilde{\gamma}_n \, \gamma_n^{1/4}.
\end{aligned}$$

The last inequality follows from the Distortion Lemma, which says that for each $s = n, n+1, \dots, 2n-2$

$$\mu^{\mathrm{st}}_{s,\tau}\{\mathcal{P}^{\mathrm{st}}_{<2n,s,\boldsymbol{X}_n}(\tau)\} \le 3.$$

Since $2^{11/4}3^{n-1} < 6^{n+1}$, this yields the required estimate (18) for $m = n-1$.

Suppose now that (18) is true for $m+1$ and we want to prove it for $m$. Denote by $G^{1,\tilde\gamma_n}_{<2n,m,\tau}(f, u(m-1); x_0, \dots, x_m) \subset I_{\tilde\gamma_n}$ the set of points $x_{m+1}$ of the $\tilde\gamma_n$-grid $I_{\tilde\gamma_n}$ such that the $(m+2)$-tuple $x_0, \dots, x_{m+1}$ is a $\tilde\gamma_n$-pseudotrajectory associated to some extension $u(m) \in \mathcal{P}^{\mathrm{st}}_{<2n,\le m,\boldsymbol{X}_m}(\tau)$ of $u(m-1)$. In other words, $G^{1,\tilde\gamma_n}_{<2n,m,\tau}(f, u(m-1); x_0, \dots, x_m)$ is the set of all possible continuations of the $\tilde\gamma_n$-pseudotrajectory $x_0, \dots, x_m$ using all possible Newton parameters $u_m$ allowed by the family (14).

Now if $x_0, \dots, x_m$ is a $\tilde\gamma_n$-pseudotrajectory associated to $u(m) = (u_0, \dots, u_m)$, then at most one value of $x_{m+1} \in I_{\tilde\gamma_n}$ are within $\tilde\gamma_n$ of $f_{u(m),\boldsymbol{X}_m}(x_m)$. Thus for fixed $u(m-1) = (u_0, \dots, u_{m-1}) \in \mathcal{P}^{\mathrm{st}}_{<2n,\le m-1,\boldsymbol{X}_n}(\tau)$, each value of $u_m \in \mathcal{P}^{\mathrm{st}}_{<2n,m,\boldsymbol{X}_n}(\tau)$ corresponds to at most one point in $G^{1,\tilde\gamma_n}_{<2n,m,\tau}(f, u(m-1); x_0, \dots, x_m)$. It follows that

$$\sum_{x_{m+1}\in G^{1,\tilde\gamma_n}_{<2n,m,\tau}(f,u(m-1);x_0,\dots,x_m)} \mu^{\mathrm{st}}_{\le m,\tau}\big\{T^{1,\tilde\gamma_n}_{<2n,\le m,\tau}(f; x_0, \dots, x_{m+1})\big\}$$
$$\le 2\,\mu^{\mathrm{st}}_{m,\tau}\{\mathcal{P}^{\mathrm{st}}_{<2n,m,\boldsymbol{X}_n}(\tau)\}\,\mu^{\mathrm{st}}_{\le m-1,\tau}\big\{T^{1,\tilde\gamma_n}_{<2n,\le m-1,\tau}(f; x_0, \dots, x_m)\big\}.$$

The Distortion Lemma then implies that

$$\sum_{x_{m+1}\in G^{1,\tilde\gamma_n}_{<2n,m,\tau}(f,u(m-1);x_0,\dots,x_m)} \mu^{\mathrm{st}}_{\le m,\tau}\big\{T^{1,\tilde\gamma_n}_{<2n,\le m,\tau}(f; x_0, \dots, x_{m+1})\big\}$$
$$\le 6\,\mu^{\mathrm{st}}_{\le m-1,\tau}\big\{T^{1,\tilde\gamma_n}_{<2n,\le m-1,\tau}(f; x_0, \dots, x_m)\big\}.$$

Inductive application of this formula completes the proof of the Collection Lemma.□

This completes an outline of treatment of part (I) of (12) for the 1-dimensional case. To carry out part (I) of (12) in the $N$-dimensional case ($N > 1$) we use the same ideas, but have to overcome additional difficulties. We discuss them in details in [KH1], sect. 4.1 (detailed discussion of the 2-dimensional case can be found in [GK], sect. 10–11 and of the $N$-dimensional case in [Ka4], sect. 8–9) and just briefly mention them here.

• (*Nonuniqueness*) It turns out that there are many ways to write Newton interpolation polynomial in $N$ variables.

• (*Dynamically essential coordinates*) Among many $N$-dimensional Newton monomials we need to choose those effective for perturbation (see [KH1] (4.6–4.7), [GK], sect. 10.1, and [Ka4], sect. 8.2).

• (*The multidimensional Distortion Lemma*) The 1-dimensional Distortion Lemma leads to an exponential factor $6^{2n}$ coming from dimension of the space of polynomials of degree $< 2n$ in 1-variable. The space of polynomials of degree $< 2n$ in $N$-variables is $\sim (2n)^N$. This forces us to find a better multidimensional Distortion Lemma (see [KH1], sect. 4.3, [Ka4], sect. 8.3, and [GK], sect. 11.4).

To treat part (II.A) of (13) we need to analyze nonsimple (recurrent) periodic trajectories of period $n$ knowing that all periodic trajectories of period $< n$ are sufficiently hyperbolic (see (I) of (12)).

## 6. Partition of nonsimple periodic trajectories into simple almost periodic parts

Analysis of nonsimple periodic trajectories of multidimensional diffeomorphisms, performed in [KH1] and [Ka4], occupies sect. 2.4 and 3.5 in [KH1] and section 5 in [Ka4]. The goal for each nonsimple periodic trajectory $\{x_j = f^j(x_0)\}_{j=0}^{n-1}$ of period $n$ find a close return, say $x_k$, so that $\{x_j\}_{j=0}^{n-1}$ nearly repeats $\{x_j\}_{j=0}^{k-1}$ exactly $n/k$ times and $\{x_j\}_{j=0}^{k-1}$ is simple. This, in particular, means that hyperbolicity of $\{x_j\}_{j=0}^{k-1}$ and $\{x_j\}_{j=0}^{n-1}$ are closely related. Here we just summarize the strategy to obtain such a partition. This is exactly the step where *we cannot handle a sequence of* $\{c\gamma_n\}_{n\geq 1}$ *that decay slower than a stretched exponential* $\exp(-n^{1+\delta})$ $(\delta > 0)$. In other words, if $\gamma_n$ decays not too fast, say exponentially, we are unable to find a close return with the above properties (see [KH1], Appendix D for further discussion).

The following definitions are the key elements of the mechanism to find a close return. They quantitatively characterize close returns.

**Definition 6.1.** Let $g$ be a diffeomorphism and let $D$ be large and positive. A point $x_0$ (resp. a trajectory $x_0, \dots, x_{n-1} = g^{n-1}(x_0)$ of length $n$) has a *weak* $(D, n)$*-gap* at a point $x_k = g^k(x_0)$ if

$$|x_k - x_0| \leq D^{-n} \min_{0<j\leq k-1} |x_0 - x_j|.$$

and there is no $m < k$ such that $x_0$ has a weak $(D, n)$-gap at $x_m = g^m(x_0)$.

This definition characterizes a close return at $x_0$. For the proof we need a modification of this definition (see [KH1], Definition 3.5.3). See the Shift Theorem [KH1], sect. 3.5 and [Ka4], sect. 5 for all the details. Recall that $\{c\gamma_n\}_{n\geq 1}$ is the sequence tracking hyperbolicity of periodic trajectories of period $n$ introduced in the beginning of Section 2.

**Definition 6.2.** Let $g$ be a $C^2$-smooth diffeomorphism. Let also $c > 0$ and $k < n$ be positive integers. We say that a point $x_0$ has a $(k, n, c)$*-leading saddle* if $|x_0 - x_k| \leq n^{-1}(c\gamma_k)^2$. Also if $x_0$ is $(n, \tilde{\gamma}_n)$-periodic, we say that $x_0$ has no $(n, c)$-leading saddles if for all $k < n$ we have that $x_0$ has no $(k, n, c)$-leading saddles.

Now start with a diffeomophism $f$ satisfying the inductive hypothesis of order $n-1$ with constants $c\Gamma$, i.e. for any $k < n$ all periodic trajectories of period $k$ are $(k, c\gamma_k)$-hyperbolic. In particular, it means that all periodic trajectories of period $k < n$ are either sinks, or sources, or saddles.

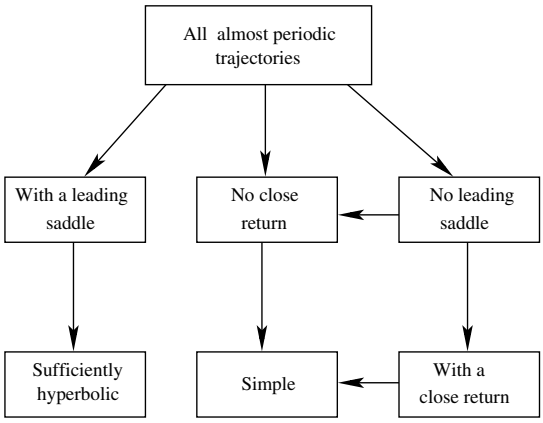



Figure 4. Various types of almost periodic periodic trajectories.

• The definition of a leading saddle is designed in such a way that if $x_0$ has a $(k, n, c)$-leading saddle, then there is a periodic point $x^* = f^k(x^*)$ close to $x_0$ (see [Ka4], Lemma 5.2.3). If $x_k, x_{2k}, \dots$ all stay close to $x^*$, then $x_0 = f^n(x_0)$ inherits hyperbolicity of $x^*$ (see [Ka4], Lemma 5.2.1).

• Suppose $x_0$ has a $(k, n, c)$-leading saddle, but for some $p < n/k$ the corresponding $x_{pk}$ leaves a small neighborhood of $x^*$. Then one can show that $x_{pk}$ has no $(n, c)$-leading saddles (see [Ka4], Lemma 5.2.4).

• Suppose $\tilde{x}_0 = x_{pk}$ has no $(n, c)$-leading saddles. It turns out that $\tilde{x}_0$ can have at most $\dim M$ weak $(D, n)$-gaps at some $\tilde{x}_{k_1}, \dots, \tilde{x}_{k_s}$, $s \le \dim M$. The reason is that each weak $(D, n)$-gap $\tilde{x}_{k_j}$ after the first one at $k_1$ implies that the linearization $df^{k_1}(\tilde{x}_0)$ has an almost eigenvalue that is a $k_j/k_1$-root of unity, and the same is true for $k_{s+1} = n$ (see [Ka4], Theorem 5.1.4).

• Suppose $\tilde{x}_0$ has no $(n, c)$-leading saddles and has $s < \dim M$ weak $(D, n)$-gaps. Then we can show that it is $(n, c\gamma_n)$-simple (see [Ka4], Theorem 5.3.1 and its extension necessary for the proof: Theorem 5.4.1).

This scheme is summarized in the diagram (see Figure 4).

## 7. Finititude of number of localized coexisting sinks

In this section we give a short exposition of a result from [GK] concerning the Newhouse phenomenon of infinitely many sinks. The primary goal of [GK] is to analyze

trajectories *localized in a neighborhood of a fixed HT*. A sink is the simplest attractor. We now introduce notions of an unfolding of a homoclinic tangency and localized trajectories of finite complexity associated to that homoclinic tangency.

Consider a 1-parameter family of perturbations $\{f_\varepsilon\}_{\varepsilon\in I}$, $I = [-\varepsilon_0, \varepsilon_0]$ of a 2-dimensional diffeomorphism $f = f_0 \in \mathrm{Diff}^r(M)$ with homoclinic tangency, where $\varepsilon_0$ is small (see Figure 5). Roughly speaking, $\varepsilon$ parameterizes oriented distance of the top tip of the unstable manifold to the stable manifold. Such a family is called an *unfolding of an HT*.

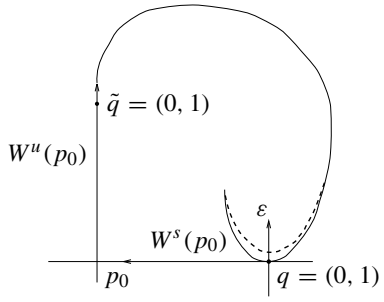



Figure 5. Homoclinic tangency.

Robinson [R], adapting Newhouse's ideas [N1], [N2], showed that for such an unfolding there is a sequence of open intervals converging to zero such that for a generic parameter from those intervals the corresponding diffeomorphism $f_\varepsilon$ has infinitely many coexisting sinks.

Assume that $f$ has a fixed saddle point $p_0 = f(p_0)$ and that the eigenvalues $\lambda$, $\mu$ of the linearization $Df(p_0)$, $0 < \lambda < 1 < \mu$, belong to the open dense set of pairs of eigenvalues for which Sternberg's linearization theorem holds. Then in a small neighborhood $\tilde{V}$ of $p_0$ there is a $C^r$ smooth normal coordinate system $(x, y) \in \tilde{V} \subset \mathbb{R}^2$ such that $f(x, y) = (\lambda x, \mu y)$. Suppose $q$ is the point of homoclinic tangency of $W^s(p_0)$ and $W^u(p_0)$ away from $\tilde{V}$, and let $\tilde{q} = f^{-1}(q)$ be its preimage.

Extend the coordinate neighborhood $\tilde{V}$ by iterating forward and backward until first it contains $\tilde{q}$ and $f(q)$, respectively. Decreasing $\tilde{V}$ if necessary we can assume that there are no overlaps. Denote such a neighborhood by $V$ and call it *a normal neighborhood*. By definition $V$ does not contain $q$ (see Figure 6). Consider a neighborhood $U$ (resp. $\tilde{U} \subset \hat{U}$) of $q$ (resp. $\tilde{q}$) such that $f(U) \cap U = \emptyset$ (resp. $f^{-1}(\hat{U}) \cap \hat{U} = \emptyset$), $f(\tilde{U}) \supset U$, and $f(\hat{U}) \cap V = \emptyset$. By rescaling coordinate axis one could set $q$ to have coordinates (1,0) and $\tilde{q}$ to have (0,1). Set $\mathcal{V} = V \cup U$. In what follows we *fix* a neighborhood $\mathcal{V}$ once and for all.

**Definition 7.1.** We call an invariant set of points $\mathcal{V}$-*localized* if it belongs to $\mathcal{V}$. In particular, any invariant set contained in

$$\Lambda_{\mathcal{V}} = \bigcap_{n\in\mathbb{Z}} f^n(\mathcal{V})$$

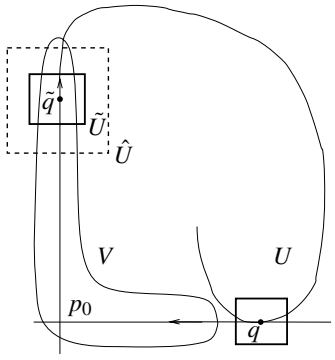



Figure 6. Localization for homoclinic tangency.

is $\mathcal{V}$-localized. A periodic point $f^n(p) = p$, $n \in \mathbb{N}$, is called $\mathcal{V}$-localized if it belongs to $\Lambda_{\mathcal{V}}$ and is called $(\mathcal{V}, s)$-localized if its trajectory $\mathfrak{P} = \{f^k(p)\}_{k=1}^n$ visits $U$ exactly $s$ times. Call $s = s(\mathfrak{P})$ the *cyclicity* of a $\mathcal{V}$-localized periodic orbit.

The zoo of $\mathcal{V}$-localized invariant sets is incredibly rich. Below we just mention the authors favorit animals: *Smale's horseshoe, infinitely many coexisting $\mathcal{V}$-localized sinks*[3], *strange attractor* (*Benedicks–Carleson* [BC], *Mora–Viana* [MV], *Young–Wang* [WY]), *arbitrarily degenerate periodic points of arbitrary high periods* (*Gonchenko–Shilnikov–Turaev* [GST1]), *uniformly and nonuniformly hyperbolic horseshoes as maximal invariant sets $\Lambda_{\mathcal{V}}$* (*Newhouse–Palis* [NP], *Palis–Takens* [PT], *Palis–Yoccoz* [PY1], [PY2], *Rios* [Ri]).

The main result of [GK] is the following

**Theorem 7.1.** *With the above notations, for a generic*[4] *1-parameter family $\{f_\varepsilon\}_{\varepsilon\in I}$ that unfolds an HT at $q$ there is a sequence of numbers $\{\boldsymbol{N}_s\}_{s\in\mathbb{N}}$ such that for almost every parameter $\varepsilon$ and any $D \in \mathbb{N}$ the corresponding $f_\varepsilon$ has only finitely many $\mathcal{V}$-localized sinks $\{\mathfrak{P}_j\}_{j\in J}$ whose cyclicity is bounded by $D$ or period exceeds $\boldsymbol{N}_{s_j}$, where $s_j = s(\mathfrak{P}_j) > D$ is cyclicity of a corresponding sink $\mathfrak{P}_j$. In other words, for almost every parameter $\varepsilon$ if there are infinitely many coexisting $\mathcal{V}$-localized sinks $\{\mathfrak{P}_j\}_{j\in J}$, then all but finitely many have cyclicity $s_j = s(\mathfrak{P}_j) > D$ and period $< \boldsymbol{N}_{s_j}$.*

**Remark 7.1.** For 1-loop periodic sinks a similar result is obtained by Tedeschini-Lalli–Yorke [LY]. Dynamical properties of periodic and homoclinic orbits of low cyclicity ($s = 1, 2, 3$) were studied in [GST1], [GStT]. In particular, Gonchenko–Shilnikov found the relation between existence of the infinite number of 2-loop sinks and numerical properties of the invariants of smooth conjugacy [GoS].

[3]Actually Newhouse [N2] (see also Palis–Takens [PT] for a simplified proof) proved that for a Baire generic set of diffeomorphisms in a Newhouse domain there are infinitely many coexisting sinks. However one can construct infinitely many of those as $\mathcal{V}$-localized.

[4]meaning of "generic" is in the sense of prevalence in the space of 1-parameter families see Section 9 for a definition.

**Remark 7.2.** We can choose $N_s = s^{5s^2}$.

Palis–Takens [PT] and Palis–Yoccoz [PY1], [PY2] investigated generic unfolding of an HT not only for saddle periodic points but also for horseshoes. They investigated parameters *outside* of Newhouse domains. We obtain less sharp properties of the dynamics, but we treat parameters *inside* Newhouse domains too.

## 8. Discussion of the proof of Theorem 7.1

To prove Theorem 7.1 we follow very similar strategy as to prove Theorem 2.1. First we introduce several notions:

**Trajectory type, hyperbolic and parabolic maps.** Any $(\mathcal{V}, s)$-localized periodic orbit, by definition, visits $U$ exactly $s$ times and spends $n_1, n_2, \dots, n_s$ consecutive iterates in $V$, $n = n_1+n_2+\ldots+n_s+s$. We call an ordered sequence $(n_1, \dots, n_s)$ *type* of a periodic orbit. For a given periodic orbit denote the points of intersection with $U$ by $\boldsymbol{p}_0,\ \boldsymbol{p}_1, \dots, \boldsymbol{p}_{s-1}$ and the corresponding points in $\tilde{U}$ by $\tilde{\boldsymbol{p}}_0 = f^{n_1}(\boldsymbol{p}_0),\ \tilde{\boldsymbol{p}}_1 = f^{n_2}(\boldsymbol{p}_1), \dots, \tilde{\boldsymbol{p}}_{s-1} = f^{n_s}(\boldsymbol{p}_{s-1})$.

Recall that $f$ is linear in $V \setminus \tilde{U}$ with eigenvalues $\lambda < 1 < \mu$, $f|_{V\setminus\tilde{U}}(x, y) = (\lambda x, \mu y)$. Call this linear map *hyperbolic*, denoted $L$, and $f|_{\tilde{U}}$ *parabolic*, denoted $\mathcal{P}$.

We replace hyperbolicity of periodic points from (1) by the *cone condition*.

**Cone condition.** To estimate the measure of parameters for which a periodic orbit of a given type is not a sink and even has exponentially large linearization, we introduce the following cone condition. Define at every point $p \in U$ a cone

$$K_A(p) = \{v = (v_x, v_y) \in T_p\mathcal{V} \simeq \mathbb{R}^2 :\ |v_y| \geq \mu^{-A}|v_x|\}.$$

To show that the periodic point $\boldsymbol{p}_0$ is hyperbolic it turns out that it suffices to find $0 < \alpha \ll 1$ independent of $n$ such that

$$Df_\varepsilon^n(K_{\alpha n}(\boldsymbol{p}_0)) \subset K_{\alpha n}(\boldsymbol{p}_0). \tag{19}$$

To verify this condition directly does not seem possible in general. Our plan is to verify that for most parameters this cone condition holds after each loop:

$$Df_\varepsilon^{n_i+1}(K_{\alpha n}(\boldsymbol{p}_i)) \ \subset\ K_{\alpha n}(\boldsymbol{p}_{i+1 \pmod s}) \quad \text{for each } i = 0, \dots, s-1. \tag{20}$$

This condition clearly implies (19), because the image of the first cone $K_{\alpha n}(\boldsymbol{p}_0)$ belongs to the second cone $K_{\alpha n}(\boldsymbol{p}_1)$. The image of the second one belongs to the third one and so on.

Fix $0 < \alpha \ll 1$. Notice that if all loops are *long*: $n_i > 3\alpha n$, then $L^{n_i}K_{\alpha n}(\boldsymbol{p}_i)$ is the cone of width angle $< 2\mu^{-\alpha n}$. Fix $1 \leq j \leq s$. To satisfy condition (20) for $j$

we need to avoid the intersection of the cone $Df_{\varepsilon,\tilde{\boldsymbol{p}}_j}(L^{n_j}K_{\alpha n}(\boldsymbol{p}_j))$ and a complement to $K_{\alpha n}(\tilde{\boldsymbol{p}}_{j+1})$ (see Figure 7 for $p = \boldsymbol{p}_{j+1}$). Assume that we can perturb $Df_{\varepsilon,\tilde{\boldsymbol{p}}_j}$ by composing with rotation and angle of rotation is a parameter. Then we need to avoid a phenomenon that has "probability" $\sim \mu^{-\alpha n}$. Taking the union over all types $\mathcal{N}_s$, $|\mathcal{N}_s| = n$ we get that probability to fail (20) for some $1 \le i \le s$ is $\sim n^s \mu^{-\alpha n}$. We avoid saying explicitly probability in what space, just assume that it is proportional to angle of rotation, and postpone the exact definition for further discussion.

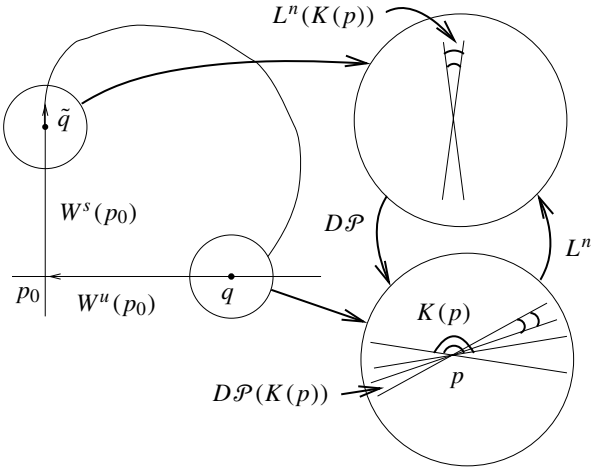



Figure 7. Evolution of cones.

However, it might happen that one of $n_i$'s is significantly smaller than $\alpha n$, e.g. $n_s \le \ln n$. In this case, the above argument fails. Indeed, let $n_s = [\ln n]$, $n \gg 1$. Consider the image of the cone $K_{\alpha n}(\boldsymbol{p}_{s-1})$ after the last loop $L^{n_s}K_{\alpha n}(\boldsymbol{p}_{s-1})$. It is the cone, whose width angle is of order 1. Taking into account possibility that $Df_{\varepsilon,\tilde{\boldsymbol{p}}_{s-1}}$ rotates a vertical vector by $\frac{\pi}{2}$ it is certainly not possible to fulfill (20) by a small perturbation. The natural idea is *to avoid looking at condition* (20) *after "short" loops*. This leads to combinatorial analysis of type $\mathcal{N}_s$ of trajectories.

**Combinatorial analysis of type $\mathcal{N}_s$ of $s$-loop trajectories.** Below we do not pay attention to dynamics of a trajectory of type $\mathcal{N}_s$ under consideration. We investigate only properties of the type $\mathcal{N}_s$ itself.

• *Short and long loops* ([GK], sect. 5.1). We shall divide an $s$-tuple $\mathcal{N}_s = (n_1, \dots, n_s)$ into two groups of *long* and *short* $n_i$'s, because they correspond to loops of a trajectory. After such a division long $n_i$'s should be *much longer* than short $n_i$'s. Denote by $t$ (resp. $s - t$) the number of long (resp. short) loops.

• *Generalized loops and essential returns* ([GK], sect. 5.2). Since we cannot fulfill (20) after short loops, we combine all loops into groups, called *generalized loops*. Each generalized loop starts with a long loop and is completed by all short loops

following afterwards. Therefore, the number of generalized loops equals the number of long loops. Then we *verify* (20) *not after each loop, but after each generalized loop*. Denote by $P_0, \dots, P_{t-1}, P_t = P_0 \subset U$ starting points of generalized loops, by $\tilde{P}_0, \dots, \tilde{P}_{t-1}, \tilde{P}_t = \tilde{P}_0$, prestarting points of generalized loops, i.e. $f(\tilde{P}_i) = P_{i+1}$, $i = 0, \dots, t-1$, and by $N_1, \dots, N_t$ their lengths respectively. Then we modify (20) to

$$Df^{N_{i+1}}(K_{\alpha n}(P_i)) \subset K_{\alpha n}(P_{i+1}) \quad \text{for each } i = 0, \dots, t-1. \tag{21}$$

Now the idea presented above has a chance to work. Indeed, let $n_j$ be a long loop and $n_{j+1}, \dots, n_{j+j'}$ be short ones from the corresponding generalized loop. Consider the image of the corresponding cone $K_{\alpha n}(P_j)$ after the generalized loop. Notice that after the long loop $n_j$ the cone $L^{n_j} K_{\alpha n}(P_j)$ is the cone of width angle $< 2\mu^{-\alpha n}$. Since long $n_j$ is so much longer than short loops $n_{j+1}, \dots, n_{j+j'}$ respectively the cone

$$\left(Df(\tilde{\boldsymbol{p}}_{j+j'}) \circ L^{n_{j+j'}} \circ \dots \circ Df(\tilde{\boldsymbol{p}}_{j+1}) \circ L^{n_{j+1}}\right) \circ \left(Df(\tilde{\boldsymbol{p}}_j) \circ L^{n_j} K_{\alpha n}(\boldsymbol{p}_j)\right)$$

has width angle $< 3\mu^{-\alpha n}$. To satisfy condition (20) for $j + j'$ we need to avoid an interval of rotations (i.e. of parameters) of length $< 5\mu^{-\alpha n}$. This phenomenon still has "probability" $\sim \mu^{-\alpha n}$.

After this combinatorial analysis we face the next difficulty. *We cannot perturb* $Df(\tilde{\boldsymbol{p}})$ *and* $Df(\tilde{\boldsymbol{p}}')$ *independently at nearby points* $\tilde{\boldsymbol{p}}$ *and* $\tilde{\boldsymbol{p}}'$.

**Dynamical analysis of trajectories.** Assume for a moment that we are interested in properties of *scattered* periodic orbits, that is, such orbits that $P_0, \dots, P_{t-1}$ in $U$ are pairwise well spaced. In particular, it is always the case for 1-loop orbits. In this case the difficulty of nearby points is removed. Using the discretization method and the cone condition (21) one can prove that for most parameters all but a finite number of the periodic orbits are hyperbolic saddles. Moreover, consider for $0 < \gamma' = \mu^{-\alpha' n} \ll \gamma'' = \mu^{-\alpha'' n}$ parameters for which a periodic not enough hyperbolic $\gamma''$-scattered $\gamma'$-pseudo-orbit of period $n$ exists. In fact, we can show that the measure of these parameters is small[5]. Now we are going to explain how this can be used to treat all periodic orbits, not necessarily scattered. Consider the 2-loop case for illustration. If starting points of loops $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are far enough from each other, one can perturb differential of parabolic map at their preimages independently, and above arguments allow to estimate the measure of "bad" parameters. Otherwise a periodic orbit can be decomposed into a union of two 1-loop periodic pseudo-orbits, which have nearby endpoints in $U$. The cone condition (21) for each of these pseudo-orbits holds for most parameters, which implies (19).

Another illustration can be given by the case $t = 1$, i.e. we have one loop which is much longer than all the others. In this case the image of the cone $K_{\alpha n}(\boldsymbol{p}_0)$ after

[5]The discretization method in this case, compare to the one described in Sections 4–5, requires certain modifications (see [GK], sect. 9–11).

the application of differential of the map along the orbit has width angle $< 2\mu^{-\alpha n}$, as explained above. Point $\tilde{\boldsymbol{p}}_{s-1} = \tilde{P}_0 = f^{n-1}(\boldsymbol{p}_0) = f^{-1}(\boldsymbol{p}_0)$ can not be too close to points $\tilde{\boldsymbol{p}}_0, \tilde{\boldsymbol{p}}_1, \ldots, \tilde{\boldsymbol{p}}_{s-2}$. Indeed, the distance between $\boldsymbol{p}_i$ and $x$-axis is $(\boldsymbol{p}_i)_y \sim \mu^{-n_{i+1}}$. Since $n_1 \gg n_i$ we have $\mu^{-n_1} \ll \mu^{-n_i}$. Therefore the point $\boldsymbol{p}_0$ can not be too close to points $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{s-1}$, and we can perturb $\phi(\tilde{\boldsymbol{p}}_{s-1}) = \phi(f^{-1}(\boldsymbol{p}_0))$ independently of $\phi(\tilde{\boldsymbol{p}}_0), \ldots, \phi(\tilde{\boldsymbol{p}}_{s-2})$. This allows to estimate the measure of "bad" parameters.

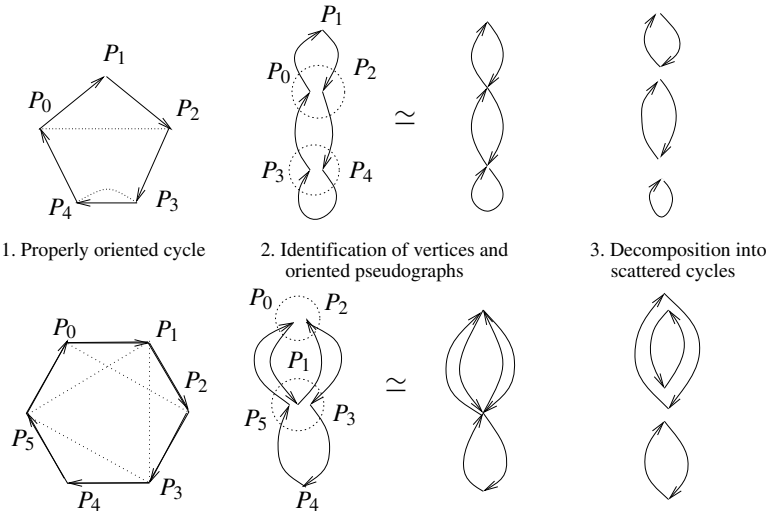



Figure 8. Graph surgery.

To consider the general case we represent a periodic orbit as an oriented cyclic graph. Starting points of generalized loops are vertices of this graph, and vertices corresponding to subsequent generalized loops are connected by an oriented edge (see Figure 8, picture 1). It turns out that for some $\gamma' \ll \gamma''$ for any pair of points $(P_i, P_j)$ either $\mathrm{dist}(P_i, P_j) > \gamma''$ or $\mathrm{dist}(P_i, P_j) < \gamma'$ (see [GK], sect. 7). Therefore every pair of vertices is either $\gamma'$-close or $\gamma''$-far apart (see Figure 8, picture 2). Now all the vertices can be divided into "clouds" or "clusters". Let us identify the vertices in the same cloud of nearby points, as shown on Figure 8, picture 2. The initial cycle is transformed now into oriented pseudograph (see [GK], Def. 20) with the same number of ingoing and outgoing edges at each vertex. Such a pseudograph can be decomposed into the union of oriented cycles (Figure 8, picture 3 and also see [GK], Lemma 7). Each of cycles from this decomposition represents a $\gamma''$-scattered $\gamma'$-pseudo-orbit. Application of the arguments above to these pseudo-orbits gives inclusion (21) for most values of parameters and implies the cone condition (19) for the initial periodic orbit.

## 9. Prevalence

Our definition of prevalence for a space $\mathrm{Diff}^r(M)$ of $C^r$ diffeomorphisms on a smooth manifold $M$ is based on the following definition from [HSY] for a complete metric linear space $V$.

**Definition 9.1** (Linear prevalence). A Borel set $S \subset V$ is called *shy* if there is a compactly supported Borel probability measure $\mu$ on $V$ such that $\mu(S - v) = 0$ for all $v \in V$. More generally, a subset of $V$ is called shy if it is contained in a shy Borel set. A subset of $V$ is called *prevalent* if its complement is shy.

(Shy sets were previously called "Haar null sets" by Christensen [Chr].) Some important properties of prevalence, proved in [HSY], are:

1. A prevalent set is dense.
2. A countable intersection of prevalent sets is prevalent.
3. A subset of $\mathbb{R}^m$ is prevalent if and only if its complement has Lebesgue measure zero.

Properties 2 and 3 above follow from the Fubini–Tonelli theorem, along with the Tychnoff theorem in the case of Property 2. Property 1 follows from the observation that a transverse measure $\mu$ can be localized in the following sense. By compactness of the support of $\mu$, there are arbitrarily small balls with positive measure. Every translation of $P$ must intersect these balls, or equivalently every translation of one of these balls must intersect $P$.

Along these lines, it is useful to think of a transverse measure for a prevalent set $P$ as a probability space of perturbations, such that at each point $v$ in the space $V$, choosing a random perturbation and adding it to $v$ yields a point in $P$ with probability one. Often the perturbations can be chosen from a finite dimensional space of parameters, using normalized Lebesgue measure on a bounded subset of parameter space. In this case, we say that $P$ is "finite-dimensionally prevalent".

In other cases, one needs an infinite number of parameters; for example, a property about periodic orbits might be finite-dimensionally prevalent for each fixed period, but higher periods require more parameters. One may be able to choose the parameters from a "Hilbert brick" $HB = J_1 \times J_2 \times \cdots$, where each $J_k$ is an interval of real numbers $\varepsilon_k$, the perturbation corresponding to $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots)$ is $\varepsilon_1 v_1 + \varepsilon_2 v_2 + \cdots$ for some vectors $v_1, v_2, \ldots \in V$, and the probability measure on $HB$ is the infinite product of normalized Lebesgue measure on each interval. We call this measure the uniform measure on $HB$. A property is then prevalent if for each $v \in V$, the property is true for $v + \varepsilon_1 v_1 + \varepsilon_2 v_2 + \cdots$ for almost every $\varepsilon$ with respect to the uniform measure on $HB$.

The notion of prevalence that we use in $\mathrm{Diff}^r(M)$ is based on this idea of choosing perturbations from a Hilbert brick. Though we cannot add perturbations in this non-linear space, by embedding $M$ in a Euclidean space $\mathbb{R}^N$, we can perturb elements of $\mathrm{Diff}^r(M)$ in a natural way by means of additive perturbations in the space $C^r(T, \mathbb{R}^N)$

of $C^r$ functions from $T$ to $\mathbb{R}^N$, where $T$ is a neighborhood of the embedded image of $M$ in $\mathbb{R}^N$. The details of this construction are given in Appendix C of [KH1]; here we provide a brief outline.

For $N$ sufficiently large, we can embed $M$ into $\mathbb{R}^N$ by the Whitney embedding theorem; choose an embedding and think of $M$ then as a subset of $\mathbb{R}^N$ (that is, identify it with its image). Choose a neighborhood $T$ of $M$ sufficiently small that the orthogonal projection $\pi : T \to M$ is well-defined. Extend each diffeomorphism $f \in \mathrm{Diff}^r(M)$ to a diffeomorphism $F$ on $T$, in such a way that $F$ is strongly contracting toward $M$. We then consider the family of perturbations

$$F_\varepsilon = F + \varepsilon_1 F_1 + \varepsilon_2 F_2 + \cdots .$$

for some functions $F_1, F_2, \ldots \in C^r(T, \mathbb{R}^N)$ and $\varepsilon$ in an appropriate Hilbert brick. For the results presented in this paper, $F_1, F_2, \ldots$ are a basis for the polynomials on $\mathbb{R}^N$, but in general they could be any functions that are chosen independently of $F$.

Next we associate to each $F_\varepsilon$ a diffeomorphism $f_\varepsilon \in \mathrm{Diff}^r(M)$. By Fenichel's theorem [Fen], for $\varepsilon$ sufficiently small, $F_\varepsilon$ has an invariant manifold $M_\varepsilon$ close to $M$, such that $\pi_\varepsilon = \pi|_{M_\varepsilon}$ is invertible. (To be precise, Fenichel's theorem is for flows, and we apply it by considering the suspension flow associated with $f$.) Furthermore, $F_\varepsilon$ is strongly contracting toward $M_\varepsilon$, so that all of its periodic orbits (indeed, all of its nonwandering points) are on $M_\varepsilon$. We then let $f_\varepsilon = \pi_\varepsilon \circ F_\varepsilon \circ \pi_\varepsilon^{-1}$. Because of this smooth conjugacy, we can prove many properties of $f_\varepsilon$ by proving them about $F_\varepsilon$.

Given this construction, we make the following definition.

**Definition 9.2** (Nonlinear prevalence). A subset $P \subset \mathrm{Diff}^r(M)$ is *prevalent* if for some functions $F_1, F_2, \ldots \in C^r(T, \mathbb{R}^N)$ and a sufficiently small Hilbert brick $HB$ such that the construction above works for every $\varepsilon \in HB$, we have that for each $f \in \mathrm{Diff}^r(M)$, the diffeomorphism $f_\varepsilon$ constructed above belongs to $P$ for almost every $\varepsilon$ with respect to the uniform measure on $HB$.

Of course, this definition depends on the choices made in our construction – the particular embedding of $M$ and the means of extending a diffeomorphism on $M$ to a neighborhood of its embedded image. We emphasize that the results in this paper and any results proved by a similar technique are independent of the details of the construction; the family of polynomial perturbations works regardless of the choices of embedding and extension. In this sense, we do not construct just a single family of perturbations for which our results are true with probability one, but rather an entire class of parametrized families that establish prevalence.

## References

[AM] Artin, M., Mazur, B., On periodic orbits. *Ann. of Math.* **81** (1965), 82–99.

[AK] Avila, A., Krikorian, R., Reducibility or non–uniform hyperbolicity for quasiperiodic Schrödinger cocycles. *Ann of Math.*, to appear.

[ALM] Avila, A., Lyubich, M., de Melo, W., Regular or stochastic dynamics in real analytic families of unimodal maps. *Invent. Math.* **154** (3) (2003), 451–550.

[BC] Benedicks, M., Carleson, L., The dynamics of the Hénon map. *Ann. of Math.* (2) **133** (1) (1991), 73–169.

[Chr] Christensen, J. P. R., On sets of Haar measure zero in abelian Polish groups. *Israel J. Math.* **13** (1972), 255–260.

[C] Colli, E., Infinitely many coexisting strange attractors. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **15** (5) (1998), 539–579.

[GaS] Gavrilov, N., Shilnikov, L., On the three dimensional system close to a system with a structurally unstable homoclinic curve. I. *Math. USSR Sb.* **17** (1972), 467–485; II. *Math. USSR Sb.* **19** (1973), 139–156.

[GoS] Gonchenko, S., Shil'nikov, L., On dynamical systems with structurally unstable homoclinic curves. *Soviet Math. Dokl.* **33** (1) (1986), 234–238.

[GST] Gonchenko, S., Shil'nikov, L., Turaev, D., Homoclinic tangencies of an arbitrary order in Newhouse regions. *Itogi Nauki Tekh. Ser. Sovrem. Mat. Prilozh. Temat. Obz., Vseross. Inst. Nauchn. i Tekhn. Inform.* **67** (1999), 69–128; English transl. *J. Math. Sci.* **105** (1) (2001), 1738–1778.

[GST1] Gonchenko, S., Shil'nikov, L., Tuvaev, D., On models with non-rough Poincare homoclinic curves. *Physica D* **62** (1–4) (1993), 1–14.

[GStT] Gonchenko, S., Sten'kin, O., Tuvaev, D., Complexity of homoclinic bifurcations and Ω-moduli. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **6** (6) (1996), 969–989.

[GK] Gorodetski, A., Kaloshin, V., How often surface diffeomorphisms have infinitely many sinks and hyperbolicity of periodic points near a homoclinic tangency. *Adv. Math.*, to appear.

[Fen] Fenichel, N., Persistence and smoothness of invariant manifolds for flows. *Indiana Univ. Math. J.* **21** (1971), 193–226.

[HSY] Hunt, B. R., Sauer, T., Yorke, J. A., Prevalence: a translation-invariant almost every for infinite dimensional spaces. *Bull. Amer. Math. Soc.* **27** (1992), 217–238; Prevalence: an addendum. *Bull. Amer. Math. Soc.* **28** (1993), 306–307.

[Ka1] Kaloshin, V., Generic diffeomorphisms with superexponential growth of the number of periodic orbits. *Comm. Math. Phys.* **211** (1) (2000), 253–271.

[Ka2] Kaloshin, V., An extension of the Artin-Mazur theorem. *Ann. of Math.* **150** (1999), 729–741.

[Ka3] Kaloshin, V., Ph.D. thesis. Stretched exponential estimate bound on the growth of the number of periodic points for prevalent diffeomorphisms. Princeton University, 2001.

[Ka4] Kaloshin, V., Stretched exponential bound on growth of the number of periodic points for prevalent diffeomorphisms, part II. Preprint, 2001; www.its. caltech.edu/ ~kaloshin/research/Per-pts2.pdf.

[Ka5] Kaloshin, V., Some prevalent properties of smooth dynamical systems. *Proc. Steklov Math. Inst.* **213** (3) (1997), 115–140.

[KH1] Kaloshin, V., Hunt, B., A stretched exponential bound on growth of the number of periodic points for prevalent diffeomorphisms, part I. *Ann. of Math.*, to appear; www.its.caltech.edu/~kaloshin/research/Per-pts1.pdf.

[KH2] Kaloshin, V., Hunt, B., A stretched exponential bound on the rate of growth of the number of periodic points for prevalent diffeomorphisms. *Electron. Res. Announc. Amer. Math. Soc.* **7** (2001), part I, 17–27; part II, 28–36.

[Ko] Kolmogorov, A. N., Théorie générale des systèmes dynamiques et mécanique classique. In *Proceedings of the International Congress of Mathematicians* (Amsterdam, 1954), Vol. 1, Erven P. Noordhoff N.V., Groningen; North-Holland Publishing Co., Amsterdam 1957, 315–333.

[Ly] Lyubich, M., Almost every real quadratic map is either regular or stochastic. *Ann. of Math* (2) **156** (1) (2002), 1, 1–78.

[MV] Mora, L., Viana, M., Abundance of strange attractors. *Acta Math.* **171** (1) (1993), 1–71.

[N1] Newhouse, S., Diffeomorphisms with infinitely many sinks. *Topology* **13** (1974), 9–18.

[N2] Newhouse, S., The abundance of wild hyperbolic sets and nonsmooth stable sets of diffeomorphisms. *Inst. Hautes Études Sci. Publ. Math.* **50** (1979), 101–151.

[NP] Newhouse, S., Palis, J., Cycles and bifurcation theory. *Asterisque* **31** (1976), 44–140.

[OY] Ott, W., Yorke, J., Prevalence. *Bull. Amer. Math. Soc.* **42** (3) (1993), 263–290.

[O] Oxtoby, J. C., *Measure and Category*. Grad. Texts in Math. 2, Springer-Verlag, Berlin 1971.

[Pa] Palis, J., A Global view of dynamics and a conjecture on the denseness of finititude of attractors. *Astérisque* **261** (2000), 335–347.

[PT] Palis, J., Takens, F., *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations*. Cambridge University Press, Cambridge 1993.

[PY1] Palis, J., Yoccoz, J.-Ch., Homoclinic tangencies for hyperbolic sets of large Hausdorff dimension. *Acta Math.* **172** (1) (1994), 91–136.

[PY2] Palis, J., Yoccoz, J.-Ch., Fers à cheval non uniformément hyperboliques engendrés par une bifurcation homocline et densité nulle des attracteurs. *C. R. Acad. Sci. Paris Sér. I Math.* **333** (9) (2001), 867–871.

[R] Robinson, C., Bifurcations to infinitely many sinks. *Comm. Math. Phys.* **90** (3) (1986), 433–459.

[Ri] Rios, I., Unfolding homoclinic tangencies inside horseshoes: hyperbolicity, fractal dimensions and persistent tangencies. *Nonlinearity* **14** (3) (2001), 431–462.

[Si] Simon, B., Operators with singular continuous spectrum. I. General operators. *Ann. of Math* (2) **141** (1) (1995), 131–145.

[LY] Tedeschini-Lalli, L., Yorke, J., How often do simple dynamical processes have infinitely many coexisting sinks? *Comm. Math. Phys.* **106** (4) (1986), 635–657.

[WY] Wang, Q., Young, L.-S., Strange attractors with one direction of instability. *Comm. Math. Phys.* **218** (1) (2001), 1–97.

California Institute of Technology, Mathematics 253-37, Pasadena, CA 91125, U.S.A.
E-mail: asgor@caltech.edu

University of Maryland, Department of Mathematics, College Park, MD 20742-3281 U.S.A.
E-mail: bhunt@ipst.umd.edu

Penn State University, Mathematics Department, University Park, State College, PA 16802, U.S.A.
and
California Institute of Technology, Mathematics 253-37, Pasadena, CA 91125, U.S.A.
E-mail: kaloshin@math.psu.edu

# From combinatorics to ergodic theory and back again

Bryna Kra*

**Abstract.** Multiple ergodic averages, such as the average of expressions like $f_1(T^n x)$ $f_2(T^{2n}x)\dots f_k(T^{kn}x)$, were first studied in the ergodic theoretic proof of Szemerédi's Theorem on arithmetic progressions. It turns out that all constraints on such averages (in a sense that we describe) have an algebraic character, arising from identities in nilpotent groups. We discuss these averages, several generalizations, and combinatorial implications of the results.

**Mathematics Subject Classification (2000).** 37A30, 11B25, 37A45.

**Keywords.** Multiple ergodic theorem, multiple recurrence, arithmetic progressions, nilsystems.

## 1. Additive combinatorics and ergodic theory

A classic result of Ramsey Theory was proved by van der Waerden [53] in the 1920s, who showed that if the integers are partitioned into finitely many subsets, at least one of the subsets contains arbitrarily long arithmetic progressions. Erdős and Turán [12] conjectured that a weaker assumption suffices: if $A$ is a set of integers whose *upper density*

$$\bar{d}(A) = \limsup_{N\to\infty} \frac{1}{N}|A \cap [1, N]|$$

is positive, then $A$ contains arbitrarily long arithmetic progressions. Clearly the conjecture immediately implies van der Waerden's Theorem.

The first progress on the Erdős–Turán conjecture came in 1952, when Roth [45] used Fourier analysis to establish that a set of integers with positive upper density contains an arithmetic progression of length 3. Further progress was not until 1969, when Szemerédi [48] showed that the conjecture holds for progressions of length 4. Finally in 1975, Szemerédi [49] resolved the general case with an intricate combinatorial proof.

Soon thereafter, Furstenberg [18] used ergodic theory to give a new proof of Szemerédi's Theorem, and this proof marks the birth of the field of ergodic Ramsey Theory. Since then, ergodic theory has been used to prove new results in combinatorics, such as the multidimensional Szemerédi Theorem [22], the density Hales–Jewett Theorem [24], and the polynomial Szemerédi Theorem [4]; many of these results have yet to be obtained by other means. (Some of these results are explained

*The author was partially supported by NSF grant DMS-0244994.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

in Section 4.) Furstenberg's pioneering work laid out the general strategy for these problems: translate the combinatorial statement into a problem on the intersection of sets in a measure preserving system and then study the average associated to this intersection. The convergence of these multiple ergodic averages is the main focus of this article. A key result is the convergence of the averages associated to Szemerédi's Theorem (see Section 2 for an explanation of the link):

**Theorem 1.1** (Host and Kra [36]). *Assume that* $(X, \mathcal{X}, \mu, T)$ *is a measure preserving system,*[1] $k \geq 1$ *is an integer, and* $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$. *Then the limit*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n} x) \ldots f_k(T^{kn} x) \tag{1}$$

*exists in* $L^2(\mu)$.

It turns out that a subsystem can be substituted for the original system without affecting the convergence or the value of the limit. Furthermore, this subsystem can be completely described algebraically, with a particular role played by nilpotent groups and their homogeneous spaces. We describe the structural analysis of measure preserving systems needed to prove this in Section 3.

This has led us to a greater understanding of other multiple ergodic averages, including averages with polynomial exponents, prime exponents, and certain averages of commuting transformations, and some of these results are discussed in Section 4. In turn, the multiple convergence theorems have lead to deeper connections with exciting developments in number theory and combinatorics, and we discuss some of these developments in Sections 4 and 5.

Although the connection between ergodic theory and additive combinatorics is well established, the depth of this connection is only now beginning to be understood. Szemerédi's original proof is combinatorial and Furstenberg's proof uses ergodic theory, yet the two proofs have many formal similarities. These features recur in more recent proofs of Szemerédi's Theorem, such as those of Gowers [26] and of Tao [50]. In the ergodic setup, with our work in [36] we have a complete understanding of the underlying structures in measure preserving systems that arise in the ergodic theoretic proof of Szemerédi's Theorem. To elucidate the true nature of the link with additive combinatorics, describing corresponding combinatorial constructions remains a deep open question.

[1] By an (*invertible*) *measure preserving* (*probability*) *system*, we mean a quadruple $(X, \mathcal{X}, \mu, T)$ where $X$ is a compact metrizable set, $\mathcal{X}$ denotes the Borel $\sigma$-algebra on $X$, $\mu$ is a probability measure on $(X, \mathcal{X})$, and $T\colon X \to X$ is an invertible measurable map with $\mu(A) = \mu(T^{-1}A)$ for all $A \in \mathcal{X}$. Even when not explicitly stated, the measure is assumed to be a probability measure and the transformation is assumed to be invertible.

## 2. Multiple ergodic averages

**2.1. Multiple recurrence.** We start with the connection between regularity properties of subsets of integers and recurrence in measure preserving systems:

**Correspondence Principle** (Furstenberg [18], [20]). *Let $E$ be a set of integers with positive upper density. There exist a measure preserving system $(X, \mathcal{X}, \mu, T)$ and a subset $A \subseteq \mathcal{X}$ such that $\mu(A) = \bar{d}(E)$ and*

$$\bar{d}((E+n_1) \cap (E+n_2) \cap \cdots \cap (E+n_k)) \geq \mu(T^{-n_1}A \cap T^{-n_2}A \cap \cdots \cap T^{-n_k}A)$$

*for any integer $k \geq 1$ and integers $n_1, n_2, \ldots, n_k \geq 0$.*

Furstenberg then deduced Szemerédi's Theorem by showing that any system $(X, \mathcal{X}, \mu, T)$ is *multiply recurrent*, meaning that for all $A \in \mathcal{X}$ with positive measure, there exists $n \in \mathbb{N}$ such that

$$\mu(A \cap T^n A \cap T^{2n} A \cap \cdots \cap T^{kn} A) > 0. \tag{2}$$

To produce such $n \in \mathbb{N}$ using ergodic theoretic methods, it is natural to average the expression in (2) over $n$. If one can show that the limit inferior of this average is positive, the existence of some $n \in \mathbb{N}$ satisfying (2) follows immediately. Thus combined with the Correspondence Principle, Szemerédi's Theorem follows from:

**Multiple Recurrence Theorem** (Furstenberg [18]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving system, $A \in \mathcal{X}$ has positive measure, and $k \geq 1$ is an integer. Then*

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^n A \cap T^{2n} A \cap \cdots \cap T^{kn} A) > 0. \tag{3}$$

Poincaré Recurrence is implied by the case $k = 1$: for any set $A \in \mathcal{X}$ with positive measure, there exist infinitely many $n \in \mathbb{N}$ such that $\mu(A \cap T^n A) > 0$. Although it is easy to prove Poincaré Recurrence directly, we can also view it as a corollary of the von Neumann Ergodic Theorem, which implies that for a set $A \in \mathcal{X}$ with positive measure, the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^n A)$$

exists and is positive. For higher order multiple recurrence ($k \geq 2$), this method of studying the corresponding multiple ergodic average is the only known method for producing $n$ such that (2) holds.

**2.2. Multiple ergodic averages.** A natural question arises: is the "lim inf" in (3) actually a limit? More generally, if $(X, \mathcal{X}, \mu, T)$ is a measure preserving system,

$k \geq 1$ is an integer, and $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$, do the multiple ergodic averages

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n}x)\ldots f_k(T^{kn}x) \tag{4}$$

converge as $N$ tends to infinity, and in what sense do they converge? Taking each $f_i$ to be the indicator function $\mathbf{1}_A$ of a set $A$, multiplying by $\mathbf{1}_A$ and integrating with respect to $\mu$, we obtain the average in (3). For $k = 1$, the existence of this limit in $L^2(\mu)$ is the von Neumann Ergodic Theorem.

A measure preserving transformation $T\colon X \to X$ induces an operator $U_T$, on functions in $L^2(\mu)$ defined by $U_T f(x) = f(Tx)$. In a standard abuse of notation, we denote the operator $U_T$ by $T$ and write $Tf(x) = f(Tx)$. In general we assume that the measure preserving system $(X, \mathcal{X}, \mu, T)$ is *ergodic*, meaning that the only sets $A \in \mathcal{X}$ satisfying $T^{-1}A \subseteq A$ have either full or zero measure. Since a general system can be decomposed into its ergodic components, for most of the theorems we consider it suffices to assume that the system is ergodic.

When the system is ergodic, for $k = 1$ the limit of (4) in $L^2(\mu)$ is the integral $\int_X f_1\, d\mu$ and in particular is constant. However, without some assumption on the system, for $k \geq 2$, the limit in (4) need not be constant. For example, if $X$ is the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, $T\colon \mathbb{T} \to \mathbb{T}$ is the rotation $Tx = x + \alpha \mod 1$ for some $\alpha \in \mathbb{T}$, $f_1(x) = \exp(4\pi i x)$ and $f_2(x) = \exp(-2\pi i x)$, then $f_1(T^n x) f_2(T^{2n}x) = f_2^{-1}(x)$ for all $n \in \mathbb{N}$. In particular, the double average

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n}x)$$

converges to a nonconstant function. (More generally, if $\alpha \notin \mathbb{Q}$ and $f_1, f_2 \in L^\infty(\mu)$, the double average converges to $\int f_1(x+t) f_2(x+2t)\, dt$, which in general is not constant.)

The limit behavior of the double average depends on rotational behavior in the system. To make this more precise, we introduce some terminology. A *factor* of a measure preserving system $(X, \mathcal{X}, \mu, T)$ can be defined in one of several equivalent ways: it is a $T$-invariant sub-$\sigma$-algebra $\mathcal{Y}$ of $\mathcal{X}$, it is a measure preserving system $(Y, \mathcal{Y}, \nu, S)$ and a measurable map $\pi\colon X \to Y$ such that $\mu \circ \pi^{-1} = \nu$ and $S \circ \pi(x) = \pi \circ T(x)$ for $\mu$-almost all $x \in X$, and it is a $T$-invariant subspace $\mathcal{F}$ of $L^\infty(\mu)$. The equivalence between the first two definitions follows by identifying $\pi^{-1}(\mathcal{Y})$ with a $T$-invariant sub-$\sigma$-algebra of $\mathcal{X}$ and noting that any $T$-invariant sub-$\sigma$-algebra of $\mathcal{X}$ arises in this way. Setting $\mathcal{F} = L^\infty(\mathcal{Y})$, we have that the first definition implies the third and taking $\mathcal{Y}$ to be the $\sigma$-algebra generated by $\mathcal{F}$-measurable sets, we have the converse. Depending on the context, we use any of these three characterizations interchangeably. In a slight abuse of notation, we use the same letter to denote the transformation in the whole space and in a factor.

If $(Y, \mathcal{Y}, \nu, T)$ is a factor of $(X, \mathcal{X}, \mu, T)$ and $f \in L^2(\mu)$, the *conditional expectation* $\mathbb{E}(f \mid \mathcal{Y})$ *of* $f$ *with respect to* $\mathcal{Y}$ is the orthogonal projection of $f$ onto $L^2(\nu)$.

Let $\mathbb{E}(f \mid Y)$ denote the function on $Y$ defined by $\mathbb{E}(f \mid Y) \circ \pi = \mathbb{E}(f \mid \mathcal{Y})$, where $\pi : X \to Y$ is the natural projection. This expectation is characterized by

$$\int_Y \mathbb{E}(f \mid Y)(y) g(y)\, d\nu(y) = \int_X f(x) g(\pi(x))\, d\mu(x)$$

for all $g \in L^\infty(\mu)$.

A measure preserving system $(X, \mathcal{X}, \mu, T)$ is said to be *weakly mixing* if the only measurable eigenfunctions of the operator on $L^2(\mu)$ induced by the transformation $T$ are constant. An alternate characterization of weakly mixing can be given in terms of a factor: the measure preserving system $(X, \mathcal{X}, \mu, T)$ is not weakly mixing if and only if it has a nontrivial factor which is a rotation on a compact abelian group. The maximal such (group rotation) factor is known as the *Kronecker factor*. A rotation on a circle is not weakly mixing.

Taking the rotational behavior into account, the double average $\frac{1}{N} \sum T^n f_1 \cdot T^{2n} f_2$ can be understood. The obvious phenomenon is that for $\mu$-almost every $x$, the triple $(x, T^n x, T^{2n} x)$ projects to an arithmetic progression in the Kronecker factor $\mathcal{Z}$. Furstenberg showed that this restriction is the only restriction, meaning that

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n f_1 \cdot T^{2n} f_2 - \frac{1}{N} \sum_{n=0}^{N-1} T^n \mathbb{E}(f_1 \mid \mathcal{Z}) \cdot T^{2n} \mathbb{E}(f_2 \mid \mathcal{Z}) \right\|_{L^2(\mu)}$$

tends to 0 as $N \to \infty$. Thus to prove convergence of the double average, it suffices to replace each $f_i$, for $i = 1, 2$, by its conditional expectation $\mathbb{E}(f_i \mid \mathcal{Z})$ on the Kronecker factor. In particular, this means that one can assume that the system is an ergodic rotation on a compact abelian group. Then one can easily use Fourier analysis to show the existence of the limit. (The Kronecker factor is said to be *characteristic* for the double average. See Section 3.1 for the general definition.) The double average is the simplest example of a "nonconventional ergodic average," where even if the system is assumed to be ergodic, the limit need not be constant.

Furthermore, if the system is assumed to be weakly mixing, Furstenberg [18] showed the existence of the limit in (4) for all $k \geq 1$. Moreover, in this case the limit takes on a particularly simple form: the average converges in $L^2(\mu)$ to the product of the integrals $\int f_1\, d\mu \int f_2\, d\mu \ldots \int f_k\, d\mu$.

For a general system, the limiting behavior for $k \geq 3$ is more complicated and group rotations do not suffice for describing the long term behavior. For example, if $f(Tx) = \lambda f(x)$ for some $|\lambda| = 1$ and $F(Tx) = f(x)F(x)$, then

$$F(T^n x) = f(x) f(Tx) \ldots f(T^{n-1} x) F(x) = \lambda^{\frac{n(n-1)}{2}} (f(x))^n F(x).$$

Therefore

$$F(x)(F(T^n x))^{-3} (F(T^{2n} x))^3 (F(T^{3n} x))^{-1} = 1.$$

Projection to the Kronecker factor does not capture the behavior of generalized eigenfunctions, meaning that there is some relation among $x$, $T^n x$, $T^{2n} x$ and $T^{3n} x$ that does

not arise from the Kronecker factor. See Furstenberg [21] for a more intricate example, showing that even such generalized eigenfunctions do not suffice in determining the limiting behavior for $k = 3$.

Using a new structural analysis for ergodic systems, we describe the algebraic constraints on $n$-tuples $x, T^n x, T^{2n} x, \ldots, T^{(k-1)n} x$, and use this to obtain convergence of the averages in (4). Existence of the limit in $L^2(\mu)$ for $k = 1$ is the von Neumann Ergodic Theorem and existence for $k = 2$ was proven by Furstenberg [18]. Existence of the limit for $k = 3$ with the hypothesis of total ergodicity, meaning that $T$ and all its powers are ergodic, was proven by Conze and Lesigne ([9], [10], and [11]); this is the first place that a natural generalization (playing a major role for higher $k$) of the Kronecker factor, a 2-step nilsystem, appears as a factor. In the general case for $k = 3$, existence was shown by Furstenberg and Weiss [25] and by Host and Kra [33] (see also [34]). We proved existence of the limit (1) for all integers $k \geq 1$ in [36] and this is the statement of Theorem 1.1. More recently, Ziegler [57] has a different approach for showing the existence of the limit in the general case. The existence of the pointwise limit is a much more difficult problem and convergence is only known for $k = 2$, due to Bourgain [8].

The key role in the analysis used to prove the existence of the limit in (1) is played by nilpotent groups and their homogeneous spaces. We start with a brief overview of the ingredients in the proof of Theorem 1.1.

## 3. Structural analysis

**3.1. Characteristic factors.** A general strategy for showing the existence of an average, such as that of (1), is to find a factor such that the limiting behavior is unchanged when each function is replaced by its conditional expectation on this factor. More precisely, a factor $\mathcal{Y} \subseteq \mathcal{X}$ is a *characteristic factor* (or more succinctly, is *characteristic*) for the average

$$\frac{1}{N} \sum_{n=0}^{N-1} T^{a_1(n)} f_1 \cdot T^{a_2(n)} f_2 \ldots T^{a_k(n)} f_k$$

if the difference between this average and the same average with each function replaced by its conditional expectation on $\mathcal{Y}$

$$\frac{1}{N} \sum_{n=0}^{N-1} T^{a_1(n)} \mathbb{E}(f_1 \mid \mathcal{Y}) \cdot T^{a_2(n)} \mathbb{E}(f_2 \mid \mathcal{Y}) \ldots T^{a_k(n)} \mathbb{E}(f_k \mid \mathcal{Y})$$

converges to 0 in $L^2(\mu)$ as $N$ tends to infinity. For example, when $a_1(n) = n$ and $a_2(n) = 2n$, the Kronecker factor is characteristic for the double average. Although the term characteristic factor only appeared explicitly in the literature fairly recently [21], the method is implicit in Furstenberg's original proof [18] of Szemerédi's Theorem.

If one can find a characteristic factor for a given average, then it suffices to prove convergence when the characteristic factor is substituted for the original system. Proving convergence for the factor is then easier when the factor has a sufficiently explicit and "simple" description.

We follow this general strategy, but with a different point of view. Rather than manipulating a particular average that we want to understand, we start with an abstract construction of characteristic factors. The construction (following [36]) is based on an inductively defined sequence of measures and of seminorms,[2] which are then used to define the factors. We now outline this construction.

**3.2. Definition of measures and seminorms.** Fix an integer $k \geq 0$. We write a point $\omega \in \{0,1\}^k$ as $\omega = \omega_1\omega_2 \dots \omega_k$ with $\omega_i \in \{0,1\}$, omitting commas and parentheses, and let $|\omega| = \omega_1 + \omega_2 + \cdots + \omega_k$. Fixing an ergodic measure preserving system $(X, \mathcal{X}, \mu, T)$, let $X^{[k]} = X^{2^k}$ and let $T^{[k]}\colon X^{[k]} \to X^{[k]}$ be the map $T \times T \times \cdots \times T$, taken $2^k$ times. Elements of $X^{[k]}$ are written $\boldsymbol{x} = (x_\omega\colon \omega \in \{0,1\}^k)$. There is a natural identification of $X^{[k+1]}$ and $X^{[k]} \times X^{[k]}$, with a point $\boldsymbol{x} \in X^{[k+1]}$ being identified with $(\boldsymbol{x}', \boldsymbol{x}'') \in X^{[k]} \times X^{[k]}$, where $x'_\omega = x_{\omega 0}$ and $x''_\omega = x_{\omega 1}$ for each $\omega \in \{0,1\}^k$.

By induction, we define a probability measure $\mu^{[k]}$ on $X^{[k]}$, that is invariant under $T^{[k]}$. Set $\mu^{[0]} = \mu$. Assume that $\mu^{[k]}$ is defined for some $k \geq 0$. Let $\mathcal{I}^{[k]}$ denote the $\sigma$-algebra of $T^{[k]}$-invariant subsets of $X^{[k]}$.

Under the natural identification of $X^{[k+1]}$ with $X^{[k]} \times X^{[k]}$, define the measure preserving (probability) system $(X^{[k+1]}, \mu^{[k+1]}, T^{[k+1]})$ to be the *relatively independent joining* of $(X^{[k]}, \mu^{[k]}, T^{[k]})$ with itself over $\mathcal{I}^{[k]}$; this means that the measure $\mu^{[k+1]}$ satisfies for all bounded functions $F'$ and $F''$ on $X^{[k]}$,

$$\int_{X^{[k+1]}} F'(\boldsymbol{x}')F''(\boldsymbol{x}')\,d\mu^{[k+1]}(x) = \int_{X^{[k]}} \mathbb{E}(F' \mid \mathcal{I}^{[k]})\,\mathbb{E}(F'' \mid \mathcal{I}^{[k]})\,d\mu^{[k]}.$$

The measure $\mu^{[k+1]}$ is invariant under $T^{[k+1]}$ and the two natural projections on $X^{[k]}$ are each $\mu^{[k]}$. By induction, each of the $2^k$ natural projections of $\mu^{[k]}$ on $X$ is equal to $\mu$. Letting $C\colon \mathbb{C} \to \mathbb{C}$ denote the conjugacy map $z \mapsto \overline{z}$, we have that for a bounded function $f$ on $X$, the integral

$$\int_{X^{[k]}} \prod_{\omega \in \{0,1\}^j} C^{|\omega|} f(x_\omega)\,d\mu^{[k]}(\boldsymbol{x})$$

is real and nonnegative.

[2] Although the definition and context are on the surface quite different, these seminorms turn out to be a generalization of the norms introduced by Gowers [26] in his proof of Szemerédi's Theorem. To recover the Gowers norms, consider the space $\mathbb{Z}/N\mathbb{Z}$, the transformation $x \mapsto x + 1 \mod N$, and the uniform measure assigning each element of $\mathbb{Z}/N\mathbb{Z}$ weight $1/N$. The Gowers norms were later used by Green and Tao [28] in a spirit closer to ergodic theory and their use in our work [36]. See [32] and [39] for more on this connection.

Therefore, for a function $f \in L^\infty(\mu)$ we can define

$$|||f|||_k = \left( \int_{X^{[k]}} \prod_{\omega \in \{0,1\}^k} C^{|\omega|} f(x_\omega)\, d\mu^{[k]}(\boldsymbol{x}) \right)^{1/2^k}.$$

One can also view this definition as an average over the cube $\{0, 1\}^k$. A convergence theorem for general averages along cubes is also proved in [36], and the connection between averages along cubes and along arithmetic progressions is more fully explained in Host [32].

Using the Ergodic Theorem and the definition of the measures, we have that for any $f \in L^\infty(\mu)$,

$$|||f|||_{k+1} = \left( \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |||f \cdot \overline{T^n f}|||_k^{2^k} \right)^{1/2^{k+1}}. \tag{5}$$

To show that the map $f \mapsto |||f|||_k$ is a seminorm on $L^\infty(\mu)$, one derives a version of the Cauchy–Schwarz inequality and uses it to show subadditivity. Positivity immediately follows from Equation (5). (See [36] for details.)

We now return to the original averages along arithmetic progressions and show that the long term behavior of the average (1) is controlled by the seminorms we have constructed:

**Theorem 3.1** (Host and Kra [36]). *Assume that $(X, \mathcal{X}, \mu, T)$ is an ergodic measure preserving probability system. Let $k \geq 1$ be an integer and assume that $f_1, f_2, \ldots, f_k$ are functions on $X$ with $\|f_1\|_\infty, \|f_2\|_\infty, \ldots, \|f_k\|_\infty \leq 1$. Then*

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n f_1 \cdot T^{2n} f_2 \ldots T^{kn} f_k \right\|_{L^2(\mu)} \leq \min_{1\leq j\leq k} (j|||f_j|||_k).$$

The proof relies on a standard method for finding characteristic factors, which is an iterated use of a variation of the van der Corput Lemma on differences (see for example [40] or [1]):

**van der Corput Lemma.** *Assume that $\mathcal{H}$ is a Hilbert space with inner product $\langle\ ,\ \rangle$ and norm $\|\cdot\|$, and that $\xi_n$, $n \geq 0$, is a sequence in $\mathcal{H}$ with $\|\xi_n\| \leq 1$ for all $n$. Then*

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \xi_n \right\|^2 \leq \limsup_{H\to\infty} \frac{1}{H} \sum_{h=0}^{H-1} \limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} \langle \xi_{n+h}, \xi_n \rangle \right|.$$

In our context, we apply this to the Hilbert space $L^2(\mu)$ of unitary operators that is naturally associated with the system $(X, \mathcal{X}, \mu, T)$. The seminorms we construct reflect $k$ successive uses of the van der Corput Lemma, with the number of steps in the iteration increasing with the complexity of the averages. Theorem 3.1 follows using induction, the Cauchy–Schwarz Inequality, and the van der Corput Lemma.

**3.3. The factors.** We then show that for every integer $k \geq 1$, the seminorms define factors $Z_{k-1}$. One presentation of these factors is obtained by defining their orthogonal complements: for $k \geq 1$, it can be shown that there exists a $T$-invariant $\sigma$-algebra $\mathcal{Z}_{k-1}$ of $\mathcal{X}$ such that for $f \in L^\infty(\mu)$,

$$\|\!| f |\!\|_k = 0 \quad \text{if and only if} \quad \mathbb{E}(f \mid \mathcal{Z}_{k-1}) = 0.$$

Therefore a bounded function $f$ is measurable with respect to $\mathcal{Z}_{k-1}$ if and only if $\int f g \, d\mu = 0$ for all functions $g \in L^\infty(\mu)$ with $\|\!| g |\!\|_{k-1} = 0$.

Then $Z_{k-1}$ is defined to be the factor of $X$ associated to the sub-$\sigma$-algebra $\mathcal{Z}_{k-1}$. Thus defined, $Z_0$ is the trivial factor, $Z_1$ is the Kronecker factor and more generally, $Z_k$ is a compact abelian group extension of $Z_{k-1}$. Furthermore, the sequence of factors is increasing

$$Z_0 \leftarrow Z_1 \leftarrow Z_2 \leftarrow \cdots \leftarrow X$$

and if $T$ is weakly mixing, then $Z_k$ is the trivial factor for all $k$. In this terminology, Theorem 3.1 states that the factor $Z_k$ is characteristic for the average (1).

The bulk of the work, and also the most technical portion, is devoted to the description of these factors. The initial idea is natural: we associate to each of these factors the group of transformations which preserves the natural cubic structure that arises in the construction. This group is nilpotent. We then conclude that for a sufficiently large (for our purposes) class of systems, this group is a Lie group and acts transitively on the space. Therefore, the constructed system is a *translation on a nilmanifold*. More precisely, if $G$ is a $k$-step nilpotent Lie group and $\Gamma$ is a discrete cocompact subgroup, then the compact space $X = G/\Gamma$ is said to be a *$k$-step nilmanifold*. The group $G$ acts on $G/\Gamma$ by left translation and the translation by a fixed element $a \in G$ is given by $T_a(g\Gamma) = (ag)\Gamma$. There exists a unique probability measure $m_{G/\Gamma}$, the *Haar measure*, on $X$ that is invariant under the action of $G$ by left translations. Fixing an element $a \in G$, we call the system $G/\Gamma$ with its associated Borel $\sigma$-algebra, Haar measure $m_{G/\Gamma}$, and translation $T_a$ a *$k$-step nilsystem*. The system $(X, \mathcal{X}, \mu, T)$ is an *inverse limit of a sequence of factors* $(X_n, \mathcal{X}_n, \mu_n, T)$ if $\mathcal{X}_n$, $n \in \mathbb{N}$, is an increasing sequence of $T$-invariant $\sigma$-algebras such that $\bigvee_{n\in\mathbb{N}} \mathcal{X}_n = \mathcal{X}$ up to a set of measure 0. If in addition each factor $(X_n, \mathcal{X}_n, \mu_n, T)$ is isomorphic to a $k$-step nilsystem for $n \in \mathbb{N}$, the system $(X, \mathcal{X}, \mu, T)$ is an *inverse limit of $k$-step nilsystems*.

The structure theorem states:

**Theorem 3.2** (Host and Kra [36]). *There exists a characteristic factor for the averages in* (1) *which is isomorphic to an inverse limit of $k$-step nilsystems.*

An expository outline of the proof is also given in Host [32]. A posteriori, the role played by the nilpotent structure is not surprising: for a $k$-step nilsystem $(X, \mathcal{X}, \mu, T)$ and $x \in X$, the $(k+1)$st term $T^k x$ of an arithmetic progression is constrained by the first $k$ terms $x, Tx, T^2x, \ldots, T^{k-1}x$.

Convergence of the linear (meaning the exponents $n, 2n, \ldots, kn$ are linear) multiple ergodic average then follows easily from general properties of nilmanifolds proved by Lesigne [43] for connected groups and proved in the general case by Leibman [41].

## 4. Generalizations of multiple convergence

**4.1. Polynomial averages.** It is natural to ask what configurations, other than arithmetic progressions, must occur in sets of integers with positive upper density. Sárközy [46] and Furstenberg [19] independently showed that if a subset of integers $E$ has positive upper density and $p\colon \mathbb{Z} \to \mathbb{Z}$ is a polynomial with $p(0) = 0$, then there exist $x, y \in E$ and $n \in \mathbb{N}$ such that $x - y = p(n)$. Furstenberg's proof used ergodic theory. Once again, Furstenberg's proof used the correspondence principle and a recurrence result, this time along polynomial times. Bergelson and Leibman generalized the recurrence result for multiple polynomials:

**Theorem 4.1** (Bergelson and Leibman [4]). *Assume that $(X, \mathcal{X}, \mu, T)$ is an invertible measure preserving system, $A \in \mathcal{X}$ has positive measure, $k \geq 1$ is an integer, and $p_1, p_2, \ldots, p_k\colon \mathbb{Z} \to \mathbb{Z}$ are polynomials with $p_j(0) = 0$ for $j = 1, 2, \ldots, n$. Then*

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-p_1(n)}A \cap T^{-p_2(n)}A \cap \cdots \cap T^{-p_k(n)}A) > 0.$$

The result in [4] is actually quite a bit stronger; they prove a multidimensional version of this statement (see Section 4.2), meaning that one replaces the $j$-th occurrence of $T$ by $T_j$, for $k$ commuting measure preserving transformations $T_1, T_2, \ldots, T_k$ of the measure space $(X, \mathcal{X}, \mu)$. A polynomial version of Szemerédi's Theorem follows immediately via Furstenberg's Correspondence Principle.

The polynomial recurrence theorem naturally leads to the corresponding convergence question for multiple polynomial averages:

**Theorem 4.2** (Host and Kra [37], Leibman [42]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving system, $k \geq 1$ is an integer, $p_1, p_2, \ldots, p_k\colon \mathbb{Z} \to \mathbb{Z}$ are polynomials, and $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$. Then the limit*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^{p_1(n)} f_1 \cdot T^{p_2(n)} f_2 \ldots T^{p_k(n)} f_k \tag{6}$$

*exists in $L^2(\mu)$.*

For a weakly mixing system, convergence of (6) was proved by Bergelson [1]. In an arbitrary measure preserving system, Furstenberg and Weiss [25] proved convergence for $k = 2$ with $p_1(n) = n$ and $p_2(n) = n^2$ and $p_1(n) = n^2$ and $p_2(n) = n^2 + n$. Weak convergence was proven in [37], as well as convergence in $L^2(\mu)$ in most cases. The remaining case, along with a generalization for multiparameter polynomials, was completed in [42].

As with the linear average corresponding to exponents $n, 2n, \ldots, kn$, the behavior of a general polynomial average is controlled by the seminorms $\|\!| \cdot |\!\|_k$. Using an inductive procedure like that of [1], the averages in (6) can be reduced to an average only

with linear exponents and we obtain a result for a polynomial average analogous to Theorem 3.1. Using the structure theorem (Theorem 3.2), we have that a characteristic factor for a polynomial average is once again an inverse limit of nilsystems.

The number of steps needed in the inductive procedure used to reduce the average (6) to linear terms depends on the choice of polynomials. As might be expected, more terms and higher degree increases the number of steps needed and so the complexity of the corresponding the nilsystem rises. However, it turns out that the linearly dependent family $\{n, 2n, \dots, kn\}$ is in some sense the most difficult. For a general polynomial family, the minimal characteristic factor $Z_k$ (meaning smallest $k$) is unknown. Yet for *rationally independent polynomials*, meaning polynomials $p_1, p_2, \dots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ with $\{1, p_1, p_2, \dots, p_k\}$ linearly independent over $\mathbb{Q}$, the characteristic factor (and therefore the value of the limit) is particularly simple and is independent of the choice of polynomials. Answering a question of Bergelson posed in [2], we show:

**Theorem 4.3** (Frantzikinakis and Kra [14]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a totally ergodic measure preserving system, $k \geq 1$ is an integer, $p_1, p_2, \dots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ are rationally independent polynomials, and $f_1, f_2, \dots, f_k \in L^\infty(\mu)$. Then*

$$\frac{1}{N}\sum_{n=0}^{N-1} T^{p_1(n)} f_1 \cdot T^{p_2(n)} f_2 \dots T^{p_k(n)} f_k - \int f_1 \, d\mu \int f_2 \, d\mu \dots \int f_k \, d\mu$$

*converges to 0 in $L^2(\mu)$ as $N \to \infty$.*

Our proof uses the machinery of the Structure Theorem, but we ultimately show that the procyclic factor (an inverse limit of cyclic groups), which is contained in the Kronecker factor, is characteristic for this average. It would be interesting to prove the theorem directly, avoiding the use of nilsystems.

**4.2. Averages for commuting transformations.** Furstenberg and Katznelson generalized multiple recurrence for commuting transformations:

**Theorem 4.4** (Furstenberg and Katznelson [22]). *Assume that $(X, \mathcal{X}, \mu)$ is a probability space, $k \geq 1$ is an integer, $T_1, T_2, \dots, T_k \colon X \to X$ are commuting measure preserving transformations, and $A \in \mathcal{X}$ has positive measure. Then*

$$\liminf_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mu(A \cap T_1^{-n} A \cap T_2^{-n} A \cap \dots \cap T_k^{-n} A) > 0.$$

(Other generalizations, including combinations of the commuting and polynomial averages, are contained in [23] and [5].) Furstenberg's correspondence principle immediately implies a combinatorial version, the multidimensional version of Szemerédi's Theorem.

Once again, it is natural to ask about convergence of the corresponding commuting multiple ergodic average:

**Question 4.5.** If $k \geq 1$ is an integer, $T_1, T_2, \ldots, T_k \colon X \to X$ are commuting measure preserving transformations of a probability space $(X, \mathcal{X}, \mu)$, $p_1, p_2, \ldots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ are polynomials, and $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$, does

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T_1^{p_1(n)} f_1 \cdot T_2^{p_2(n)} f_2 \ldots T_k^{p_k(n)} f_k$$

exist in $L^2(\mu)$?

For two transformations and exponents $p_1(n) = p_2(n) = n$, existence of the limit in $L^2(\mu)$ was shown by Conze and Lesigne [9]. For arbitrary $k \geq 1$, under the assumptions that $T_j$ is ergodic for $j = 1, 2, \ldots, k$ and that $T_i T_j^{-1}$ is ergodic for $i \neq j$, $i, j \in \{1, 2, \ldots, k\}$, existence of the limit with exponents $p_1(n) = p_2(n) = \cdots = p_k(n) = n$ in $L^2(\mu)$ is shown in [15]. However, the general case (even with all exponents equal to $n$) remains open and it is easy to construct systems such that the characteristic factors are not nilsystems.

**4.3. Sequences related to prime numbers.** Recently, a new chapter in ergodic Ramsey Theory was opened, with ergodic theoretic techniques adapted for use outside of the field. A particularly spectacular result in this direction is Green and Tao's proof [28] that the prime numbers contain arbitrarily long arithmetic progressions. The connections between the proof of Green and Tao and ergodic theory are further explained in the expository articles of Host [32], Kra [39], and Tao [51]. In turn, Green and Tao's results make it possible to study convergence for other multiple ergodic averages, leading us to a greater understanding of patterns in a set of integers with positive upper density. Green and Tao ([29], [30], [31]) proved a strong uniformity result on the prime numbers and using this result, we can show that the shifted primes have multiple recurrence properties. (See also the survey articles of Green [27] and of Tao [52].) Letting $\mathbb{P}$ denote the primes, we show:

**Theorem 4.6** (Frantzikinakis, Host, and Kra [13]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving system and $A \in \mathcal{X}$ has positive measure. Then there exists $n \in \mathbb{P} - 1$ such that*

$$\mu(A \cap T^{-n} A \cap T^{-2n} A) > 0.$$

The same statement holds with $\mathbb{P} - 1$ replaced by $\mathbb{P} + 1$. Thus the shifted primes $\mathbb{P} - 1$ and $\mathbb{P} + 1$ are sets of 2-recurrence. For single recurrence, this was proven by Sárközy [47] and reproved using ergodic methods by Wierdl [54]. (Bourgain [7] and Wierdl [55] also proved several stronger results on pointwise convergence along primes.) An immediate corollary of Theorem 4.6 is that a set of integers with positive upper density contains infinitely many arithmetic progressions of length 3 whose common difference is of the form $p - 1$ for some prime $p$ (and similarly of the form $p + 1$).

Roughly speaking, we prove this by comparing the associated double average along primes with the usual double average, and show that the difference converges to 0. This relies on the uniformity result on the prime numbers of Green and Tao. It turns out that the Kronecker factor is characteristic for the associated average. The added complication is that one must work with $\mathbb{Z}/N\mathbb{Z}$ as the underlying space instead of $\mathbb{Z}$.

Using the same methods, we also show the existence of the related double ergodic average:

**Theorem 4.7** (Frantzikinakis, Host, and Kra [13]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving system and $f_1, f_2 \in L^\infty(\mu)$. Then*

$$\lim_{N\to\infty} \frac{1}{|\mathbb{P} \cap [0, N)|} \sum_{n\in\mathbb{P}, n<N} T^n f_1 \cdot T^{2n} f_2$$

*exists in $L^2(\mu)$.*

The same reduction to a uniformity statement about the prime numbers, for both recurrence and convergence, works for multiple recurrence and convergence of all lengths. However, the needed uniformity result for prime numbers remains open for longer progressions.

## 5. Lower bounds for multiple ergodic averages

**5.1. Khintchine Recurrence.** As described in Section 2, the first step in Furstenberg's Multiple Recurrence Theorem (Poincaré Recurrence) is an immediate corollary of the von Neumann Ergodic Theorem. However, using the full description of the limit, and not only positivity of the limit inferior, one can make a finer statement about the frequency of recurrence. More precisely, a set $E \subseteq \mathbb{Z}$ is *syndetic*[3] if there exists $M \in \mathbb{N}$ such that every interval of length $M$ has nontrivial intersection with the set $E$. Khintchine generalized Poincaré Recurrence and showed:

**Theorem 5.1** (Khintchine [38]). *If $(X, \mathcal{X}, \mu, T)$ is a measure preserving system and $A \in \mathcal{X}$, then for all $\varepsilon > 0$, the set*

$$\{n \in \mathbb{Z} \colon \mu(A \cap T^n A) > \mu(A)^2 - \varepsilon\}$$

*is syndetic.*

As this result follows easily from the von Neumann Ergodic Theorem, one can ask for the analogous results corresponding to other multiple recurrence theorems: if $(X, \mathcal{X}, \mu, T)$ is a measure preserving system, $A \in \mathcal{X}$, $k \geq 1$ is an integer,

[3] A syndetic set is sometimes known in the literature as *relatively dense*. A syndetic set in $\mathbb{Z}$ is said to have *bounded gaps*.

$p_1, p_2, \dots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ are polynomials with $p_j(0) = 0$ for $j = 1, 2, \dots, k$, and $\varepsilon > 0$, is the set

$$\{n \in \mathbb{Z} \colon \mu(A \cap T^{-p_1(n)} A \cap \cdots \cap T^{p_k(n)} A) > \mu(A)^{k+1} - \varepsilon\} \tag{7}$$

syndetic?

Surprisingly enough, the answer depends on the number $k$ of polynomials and on the linear dependencies among the polynomials. For rationally independent polynomials, using the fact that a characteristic factor takes on a simple form, we show that the measure of the intersection in (7) is as large as possible on a syndetic set:

**Theorem 5.2** (Frantzikinakis and Kra [16]). *Assume that $(X, \mathcal{X}, \mu, T)$ is an invertible measure preserving system, $A \in \mathcal{X}$, $k \geq 1$ is an integer, and $p_1, p_2, \dots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ are rationally independent polynomials with $p_j(0) = 0$ for $j = 1, 2, \dots, k$. Then for all $\varepsilon > 0$, the set*

$$\{n \in \mathbb{Z} \colon \mu(A \cap T^{p_1(n)} A \cap T^{p_2(n)} A \cap \cdots \cap T^{p_k(n)} A) \geq \mu(A)^{k+1} - \varepsilon\}$$

*is syndetic.*

This sharply contrasts the behavior for a family of linearly dependent polynomials, such as the linear polynomials corresponding to Szemerédi's Theorem, where the behavior depends on the number of linear terms. This dependence is illustrated in the following two theorems:

**Theorem 5.3** (Bergelson, Host, and Kra [3]). *Assume that $(X, \mathcal{X}, \mu, T)$ is an ergodic measure preserving system, $A \in \mathcal{X}$, and $k \geq 1$ is an integer. Then for all $\varepsilon > 0$, the sets*

$$\{n \in \mathbb{Z} \colon \mu(A \cap T^n A \cap T^{2n} A) \geq \mu(A)^3 - \varepsilon\}$$

*and*

$$\{n \in \mathbb{Z} \colon \mu(A \cap T^n A \cap T^{2n} A \cap T^{3n} A) \geq \mu(A)^4 - \varepsilon\}$$

*are syndetic.*

While ergodicity is not needed in Khintchine's Theorem, it is a necessary hypothesis in Theorem 5.3. In [3], we construct a counterexample for the nonergodic case.

For arithmetic progressions of length $\geq 5$, the analogous result does not hold. Based on a result of Ruzsa contained in the Appendix of [3], we show

**Theorem 5.4** (Bergelson, Host, and Kra [3]). *There exists an ergodic system $(X, \mathcal{X}, \mu, T)$ such that for all integers $\ell \geq 1$ and all $\varepsilon > 0$, there exists a set $A = A(\ell, \varepsilon) \in \mathcal{X}$ with positive measure such that*

$$\mu(A \cap T^n A \cap T^{2n} A \cap T^{3n} A \cap T^{4n} A) \leq \varepsilon \mu(A)^\ell$$

*for every integer $n \neq 0$.*

The proofs of these theorems are based on a decomposition result for the *multicorrelation sequence*

$$\int f(x) f(T^n x) f(T^{2n} x) \dots f(T^{kn} x)\, d\mu(x), \tag{8}$$

where $(X, \mathcal{X}, \mu, T)$ is a measure preserving system, $f \in L^\infty(\mu)$, and $k, n \geq 1$ are integers. We decompose such a sequence into two pieces, one of which is small in terms of density and the second of which arises from a nilsystem. We need some terminology to describe this decomposition more precisely. A bounded sequence $\{a_n\}_{n\in\mathbb{Z}}$ *tends to zero in uniform density* if

$$\lim_{N\to\infty} \sup_{M\in\mathbb{Z}} \frac{1}{N} \sum_{n=M}^{M+N-1} |a_n| = 0.$$

If $k \geq 1$ is an integer, the sequence $\{x_n\}$ is said to be a *basic $k$-step nilsequence* if there exists some $k$-step nilmanifold $X = G/\Gamma$, a continuous real valued function $\phi$ on $X$, $a \in G$ and $e \in X$ such that $x_n = \phi(a^n \cdot e)$ for all $n \in \mathbb{N}$. A *$k$-step nilsequence* is a uniform limit of basic $k$-step nilsequences. The general decomposition result is:

**Theorem 5.5** (Bergelson, Host, and Kra [3]). *Assume that $(X, \mathcal{X}, \mu, T)$ is an ergodic measure preserving system, $k \geq 1$ is an integer, and $f \in L^\infty(\mu)$. The multicorrelation sequence* (8) *is the sum of a sequence tending to zero in uniform density and a $k$-step nilsequence.*

By subtracting a sequence of integers that tends to 0 in uniform density, the sequences in Theorem 5.3 have the same behavior as the associated nilsequences (of lengths 3 and 4), and the problem reduces to studying lower bounds for the associated nilsequences. The factors constructed in [36] are used to understand the structure of these nilsequences and a key ingredient comes from the explicit formula for the average (1) given in Ziegler [56] (an alternate proof is given in [3]).

In [16], we prove a similar multicorrelation result for independent polynomials. In this case, the nilsequence takes on a simple form, as it is induced by a unipotent affine transformation.

**5.2. Combinatorial Implications.** Via a small modification of Furstenberg's Correspondence Principle, each of these results translates to a combinatorial result. The *upper Banach density* $d^*(E)$ of a set $E \subseteq \mathbb{Z}$ is defined by

$$d^*(E) = \lim_{N\to+\infty} \sup_{M\in\mathbb{Z}} \frac{1}{N} |E \cap [M, M+N-1]|.$$

Let $\varepsilon > 0$, $E \subseteq \mathbb{Z}$ have positive upper Banach density, and consider the set

$$\big\{n \in \mathbb{Z}\colon d^*\big(E \cap (E + p_1(n)) \cap \dots \cap (E + p_k(n))\big) \geq d^*(E)^{k+1} - \varepsilon\big\}. \tag{9}$$

For $k = 2$ or $k = 3$ and $p_j(n) = jn$ for $j = 1, 2, 3$, this set is syndetic, while for $k \geq 4$ and $p_j(n) = jn$ for $j = 1, 2, \ldots, k$, there exists a set of integers $E$ with positive upper Banach density such that the set in (9) is empty. On the other hand, in [16] we show that for all integers $k \geq 1$, if $p_1, p_2, \ldots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ are rationally independent polynomials with $p_i(0) = 0$ for $i = 1, 2, \ldots, k$, then the set in (9) is always syndetic.

**Question 5.6.** If $\varepsilon > 0$, $E \subseteq \mathbb{Z}$ has positive upper Banach density, for which polynomials $p_1, p_2, \ldots, p_k \colon \mathbb{Z} \to \mathbb{Z}$ with $p_i(0) = 0$ for $i = 1, 2, \ldots, k$, is the set

$$\left\{n \in \mathbb{Z} \colon d^*\big(E \cap (E + p_1(n)) \cap (E + p_2(n)) \cap \cdots \cap (E + p_k(n))\big) \geq d^*(E)^{k+1} - \varepsilon\right\}$$

syndetic?

For the polynomials of Theorems 5.2, 5.3, and 5.4 we know the answer and it is sometimes yes and sometimes no. It is reasonable to conjecture that the answer is yes for $k = 2$ and 3, as we know it holds for two extreme cases: 2 (or 3) rationally independent polynomials and 2 (or 3) linear polynomials. For higher $k$, it may be possible to lift the independence condition of Theorem 5.2 under certain circumstances. The natural approach to the problem is via the corresponding statement in ergodic theory. A first step in answering this question is finding a general formula for the multiple polynomial average (6), generalizing the formula for the linear average (1).

## 6. Future directions

**6.1. Convergence along other sequences.** General conditions on sequences of integers under which one can prove a multiple ergodic theorem are unknown:

**Question 6.1.** If $k \geq 1$ is an integer and $a_1(n), a_2(n), \ldots, a_k(n)$ are sequences of integers with $a_j(n) \to \infty$ as $n \to \infty$ for $j = 1, 2, \ldots, n$, when does

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^{a_1(n)} f_1 \cdot T^{a_2(n)} f_1 \ldots T^{a_k(n)} f_k$$

exist in $L^2(\mu)$ for all measure preserving systems $(X, \mathcal{X}, \mu, T)$ and $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$?

For $k = 1$, convenient necessary and sufficient conditions are given by the spectral theorem. However for $k \geq 2$, there is no such characterization and the proofs of multiple convergence for all known sequences (including arithmetic progressions, polynomials, and the primes) rely in some manner on a use of the van der Corput Lemma. Finding alternate proofs not relying on the van der Corput Lemma is a first step in describing choices for the sequences $a_j(n)$; a full characterization would probably require some sort of higher order spectral theorem.

Another natural question is the convergence of random multiple ergodic averages. Let $(\Omega, \mathcal{B}, P)$ be a probability space and let $\{Y_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables taking on values 0 and 1. Given $\omega \in \Omega$, let $E = E(\omega) = \{n \in \mathbb{N}\colon Y_n(\omega) = 1\}$. Ordering $E$ by size, we have defined a *random sequence* $\{a(n) = a(n, \omega)\}$ of natural numbers.

**Question 6.2.** Assume that $k \geq 1$ is an integer and that $a(n)$ is a random sequence of natural numbers generated by a sequence of independent random variables on some probability space $(\Omega, \mathcal{B}, P)$. When does

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^{a(n)} f_1 \cdot T^{2a(n)} f_2 \dots T^{ka(n)} f_k$$

exist in $L^2(\mu)$ for all measure preserving systems $(X, \mathcal{X}, \mu, T)$ and $f_1, f_2, \dots, f_k \in L^\infty(\mu)$?

For $k = 1$, Bourgain [6] showed that for a random nonlacunary sequence, meaning a sequence where $P(X_n(\omega) = 1) = p_n$ satisfies $\lim_{n\to\infty} np_n = \infty$, which is also decreasing, one has convergence in $L^2(\mu)$. For $k \geq 1$, convergence for $k = 1$ is of course a necessary condition, but it is not know if this is sufficient.

**6.2. Connections with additive combinatorics.** Since Furstenberg's proof of Szemerédi's Theorem, there has been a long and fruitful interaction between additive combinatorics and ergodic theory, with results and techniques passing from one field to the other. A major challenge remains: understand the mathematics behind the deep analogies between the two fields. The nilsystems that arise in the structural analysis of measure preserving systems should have some sort of combinatorial analog:

**Question 6.3.** What is the combinatorial analog of the Structure Theorem (Theorem 3.2)?

The uniformity norms on $\mathbb{Z}/N\mathbb{Z}$ (used in Gowers's [26] proof of Szemerédi's Theorem and in Green and Tao's [28] proof that the primes contain arbitrarily long progressions) play a role similar to the role that the seminorms described in Section 3 play in proving convergence of the multiple ergodic average along arithmetic progressions in [36]. A partial answer to this question is given by Green and Tao in [29], in which they show that generalized quadratic functions control the third uniformity norm, analogous to the way that 2-step nilsystems control the third seminorm. These generalized quadratic functions are controlled by 2-step nilsequences, and this gives a partial understanding of the combinatorial objects. It should be interesting and useful to obtain a more complete understanding of the precise nature of the link between these generalized quadratic functions and 2-step nilsequences, with a description in the finite setting of $\mathbb{Z}/N\mathbb{Z}$ rather than in $\mathbb{Z}$. For longer progressions, even partial results are not known.

It is not clear if one can directly use ergodic theory to prove statements about the primes, as the primes have zero density and Furstenberg's Correspondence Principle only applies for sets of positive upper density. However, even without a version of the Correspondence Principle that applies to zero density subsets, ergodic theory and especially its techniques has and will be further used to understand the finer structure of the primes. In analogy with multiple ergodic averages along polynomial sequences (and the use of seminorms), one may hope to combine techniques of additive combinatorics and ergodic theory to show, for example, that for all integers $k > 1$, there exist infinitely many pairs $(p, n)$ of integers with $p, n \geq 1$ such that $p$, $p + n$, $p + n^2, \ldots, p + n^k$ consists only of prime numbers.

## References


[1] Bergelson, V., Weakly mixing PET. *Ergodic Theory Dynam. Systems* **7** (1987), 337–349.

[2] Bergelson, V., Ergodic Ramsey theory an update. In *Ergodic Theory of $\mathbb{Z}^d$-actions* (ed. by M. Pollicott, K. Schmidt), London Math. Soc. Lecture Note Ser. 228, Cambridge University Press, Cambridge 1996, 1–61.

[3] Bergelson, V., Host, B., and Kra, B. (with an Appendix by I. Ruzsa), Multiple recurrence and nilsequences. *Invent. Math.* **160** (2005), 261–303.

[4] Bergelson, V., and Leibman, A., Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.* **9** (1996), 725–753.

[5] Bergelson, V., and McCutcheon, R., An Ergodic IP Polynomial Szemerédi Theorem. *Mem. Amer. Mat. Soc.* **146**, 2000.

[6] Bourgain, J., On the maximal ergodic theorem for certain subsets of the positive integers. *Israel J. Math.* **61** (1988), 39–72.

[7] Bourgain, J., An approach to pointwise ergodic theorems. In *Geometric aspects of functional analysis* (1986/87), Lecture Notes in Math. 1317, Springer-Verlag, Berlin 1988, 204–223.

[8] Bourgain, J., Pointwise ergodic theorems for arithmetic sets. *Inst. Hautes Études Sci. Publ. Math.* **69** (1989), 5–45.

[9] Conze, J.-P., and Lesigne, E., Théorèmes ergodiques pour des mesures diagonales. *Bull. Soc. Math. France* **112** (1984), 143–175.

[10] Conze, J.-P., and Lesigne, E., Sur un théorème ergodique pour des mesures diagonales. *Publ. Inst. Rech. Math. Rennes* **1** (1987), 1–31.

[11] Conze, J.-P., and Lesigne, E., Sur un théorème ergodique pour des mesures diagonales. *C. R. Acad. Sci. Paris Sér. I Math.* **306** (1988), 491–493.

[12] Erdős, P., and Turán, P., On some sequences of integers. *J. London Math. Soc.* **11** (1936), 261–264.

[13] Frantzikinakis, N., Host, B., and Kra, B., Multiple recurrence and convergence for sequences related to the prime numbers. Preprint, 2005.

[14] Frantzikinakis, N., and Kra, B., Polynomial averages converge to the product of the integrals. *Israel J. Math.* **148** (2005), 267–276.

[15] Frantzikinakis, N., and Kra, B., Convergence of multiple ergodic averages for some commuting transformations. *Ergodic Theory Dynam. Systems* **25** (2005), 799–809.

[16] Frantzikinakis, N., and Kra, B., Ergodic averages for independent polynomials and applications. *J. London Math. Soc.*, to appear.

[17] Furstenberg, H., Strict ergodicity and transformations of the torus. *Amer. J. Math.* **83** (1961), 573–601.

[18] Furstenberg, H., Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Analyse Math.* **31** (1977), 204–256.

[19] Furstenberg, H., Poincaré recurrence and number theory. *Bull. Amer. Math. Soc.* **5** (1981), 211–234.

[20] Furstenberg, H., *Recurrence in Ergodic Theory and Combinatorial Number Theory*. M. B. Porter Lectures, Princeton University Press, Princeton, N.J., 1981.

[21] Furstenberg, H., Nonconventional ergodic averages. *Proc. Sympos. Pure Math.* **50** (1990), 43–56.

[22] Furstenberg, H., and Katznelson, Y., An ergodic Szemerédi theorem for commuting transformation. *J. Analyse Math.* **34** (1979), 275–291.

[23] Furstenberg, H., and Katznelson, Y., An ergodic Szemerédi theorem for IP-systems and combinatorial theory. *J. Analyse Math.* **45** (1985), 117–268.

[24] Furstenberg, H., and Katznelson, Y., A density version of the Hales-Jewett theorem. *J. Analyse Math.* **57** (1991), 64–119.

[25] Furstenberg, H., and Weiss, B., A mean ergodic theorem for $\frac{1}{N}\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. In *Convergence in Ergodic Theory and Probability* (ed. by V. Bergelson, P. March, J. Rosenblatt), Ohio State Univ. Math. Res. Inst. Publ. 5, Walter de Gruyter, Berlin, New York 1996, 193–227

[26] Gowers., A new proof of Szemerédi's theorem. *Geom. Funct. Anal.* **11** (2001), 465–588; Erratum ibid. **11** (2001), 869.

[27] Green, B., Generalising the Hardy–Littlewood method for primes. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 373–399.

[28] Green, B., and Tao, T., The primes contain arbitrarily long arithmetic progressions. *Ann. of Math.*, to appear.

[29] Green, B., and Tao, T., An inverse theorem for the Gowers $U^3$ norm. Preprint, 2005.

[30] Green, B., and Tao, T., Quadratic uniformity of the Möbius function. Preprint, 2005.

[31] Green, B., and Tao, T., Linear equations in primes. Preprint, 2006.

[32] Host, B., Convergence of multiple ergodic averages. In *Proceedings of School on Information and Randomness*, Universidad de Chile, Santiago, Chile, to appear.

[33] Host, B., and Kra, B., Convergence of Conze-Lesigne averages. *Ergodic Theory Dynam. Systems* **21** (2001), 493–509.

[34] Host, B., and Kra, B., An odd Furstenberg-Szemerédi theorem and quasi-affine systems. *J. Analyse Math.* **86** (2002), 183–220.

[35] Host, B., and Kra, B., Averaging along cubes. In *Modern dynamical systems and applications*, Cambridge University Press, Cambridge 2004, 123–144.

[36] Host, B., and Kra, B., Nonconventional ergodic averages and nilmanifolds. *Ann. of Math.* **161** (2005), 397–488.

[37] Host, B., and Kra, B., Convergence of polynomial ergodic averages. *Israel J. Math.* **149** (2005), 1–19.

[38] Khintchine, A. Y., Eine Verschärfung des Poincaréschen "Wiederkehrsatzes". *Comp. Math.* **1** (1934), 177–179.

[39] Kra, B., The Green-Tao Theorem on arithmetic progressions in the primes: an ergodic point of view. *Bull. Amer. Math. Soc.* **43** (2006), 3–23.

[40] Kuipers, L., and Niederreiter, N., *Uniform distribution of sequences*. Pure and Applied Mathematics, Wiley & Sons, New York 1974.

[41] Leibman, A., Pointwise convergence of ergodic averages for polynomial sequences of translations on a nilmanifold. *Ergodic Theory Dynam. Systems* **25** (2005), 201–213.

[42] Leibman, A., Convergence of multiple ergodic averages along polynomials of several variables. *Israel J. Math.* **146** (2005), 303–316.

[43] Lesigne, E., Sur une nil-variété, les parties minimales associeèe á une translation sont uniquement ergodiques. *Ergodic Theory Dynam. Systems* **11** (1991), 379–391.

[44] Parry, W., Ergodic properties of affine transformations and flows on nilmanifolds. *Amer. J. Math.* **91** (1969), 757–771.

[45] Roth, K., Sur quelques ensembles d'entiers. *C. R. Acad. Sci. Paris* **234** (1952), 388–390.

[46] Sárközy, A., On difference sets of integers I. *Acta Math. Acad. Sci. Hungar.* **31** (1978), 125–149.

[47] Sárközy, A., On difference sets of integers III. *Acta Math. Acad. Sci. Hungar.* **31** (1978), 355–386.

[48] Szemerédi, E., On sets of integers containing no four elements in arithmetic progression. *Acta Math. Acad. Sci. Hungar.* **20** (1969), 89–104.

[49] Szemerédi, E., On sets of integers containing no $k$ elements in arithmetic progression. *Acta Arith.* **27** (1975), 199–245.

[50] Tao, T., A quantitative ergodic theory proof of Szemerédi's theorem. Preprint, 2004.

[51] Tao, T., Obstructions to uniformity, and arithmetic patterns in the primes. Preprint, 2005.

[52] Tao, T., The dichotomy between structure and randomness, arithmetic progressions, and the primes. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2006/2007.

[53] van der Waerden, B. L., Beweis einer Baudetschen Vermutung. *Nieuw Arch. Wisk.* **15** (1927), 212–216.

[54] Wierdl, M., Almost everywhere convergence and recurrence along subsequences in ergodic theory. PhD Thesis, Ohio State University, 1989.

[55] Wierdl, M., Pointwise ergodic theorem along the prime numbers. *Israel J. Math.* **64** (1988), 315–336.

[56] Ziegler, T., A non-conventional ergodic theorem for a nilsystem. *Ergodic Theory Dynam. Systems* **25** (2005), 1357–1370.

[57] Ziegler, T., Universal Characteristic Factors and Furstenberg Averages. *J. Amer. Math. Soc.*, to appear .

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208-2730, U.S.A.

E-mail: kra@math.northwestern.edu

# From Brouwer theory to the study of homeomorphisms of surfaces

Patrice Le Calvez

**Abstract.** We will state an equivariant foliated version of the classical Brouwer Plane Translation Theorem and will explain how to apply this result to the study of homeomorphisms of surfaces. In particular we will explain why a diffeomorphism of a closed oriented surface of genus $\geq 1$ that is the time-one map of a time dependent Hamiltonian vector field has infinitely many periodic orbits. This gives a positive answer in the case of surfaces to a more general question stated by C. Conley. We will give a survey of some recent results on homeomorphisms and diffeomorphisms of surfaces and will explain the links with the improved version of Brouwer's theorem.

**Mathematics Subject Classification (2000).** Primary 37E30, 37E35, 37J10; Secondary 37E45.

**Keywords.** Brouwer homeomorphism, Hamiltonian homeomorphism, periodic point, foliation on a surface, rotation number.

## 1. Introduction

It is a natural problem to ask if a given dynamical statement about time independent vector fields may be extended to periodic time dependent vector fields. Recall that a complete periodic time dependent smooth vector field on a manifold $M$ defines a family $(F_t)_{t\in\mathbb{R}}$ of diffeomorphisms such that $F_0 = \mathrm{Id}_M$ and $F_{t+T} = F_t \circ F_T$, for every $t \in \mathbb{R}$, if $T$ is the period. To study this system, one usually studies the discrete dynamical system induced by $F = F_T$.

Let us begin with a very simple example. Suppose that $M$ is compact and write $\chi(M)$ for the Euler characteristic of $M$. If $F$ is a homeomorphism homotopic to the identity with a finite number of fixed points, one knows by the Lefschetz formula that

$$\sum_{F(z)=z} i(F, z) = \chi(M),$$

where $i(F, z)$ is the Lefschetz index. If $F$ is the time-one map of a flow induced by a vector field, the fixed points are necessarily the singularities of $\xi$, and the previous formula may be deduced from the Poincaré–Hopf formula

$$\sum_{\xi(z)=0} i(\xi, z) = \chi(M),$$

where $i(\xi, z)$ is the Poincaré index.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Let us now give a more interesting example. Consider a symplectic compact manifold $(M, \omega)$ and write $n_M$ (resp. $n'_M$) for the minimum number of critical points that any smooth function (resp. Morse function) defined on $M$ must have. If $F$ is the time-one map of a family $(F_t)_{t\in\mathbb{R}}$ defined by a 1-periodic time dependent Hamiltonian vector field, we say that a fixed point $z$ is *contractible* if the *trajectory* $\gamma_z\colon t \mapsto F_t(z)$, defined on $[0, 1]$, is a loop homotopic to a point. Arnold's conjecture [1] states that the number of contractible fixed points is minimized by $n_M$, and that it is minimized by $n'_M$ if every fixed point of $F$ is non-degenerate (such results are obviously true in the case where the vector field is time independent). Now the the minoration by the sum of the Betti numbers is known to be true in the non-degenerate case (Liu, Tian [49], Fukaya, Ono [31]). See [36] for a history of this problem whose first proven case ($M = \mathbb{T}^{2n} = \mathbb{R}^{2n}/\mathbb{Z}^{2n}$) was solved by Conley and Zehnder [15].

Conley conjectured that the number of contractible periodic points is infinite in the case of a torus $\mathbb{T}^{2n}$. The conjecture is true if $F$ is a diffeomorphism with no degenerate fixed points (Salamon–Zehnder [53]). We will explain in Section 5 why Conley's conjecture is true if $M = \mathbb{T}^2$ and, more generally, if $M$ is a closed surface of genus $\geq 1$. The key result is an equivariant foliated version of the Brouwer Plane Translation Theorem, which will be stated in Section 4. Roughly speaking, it asserts that if $(F_t)_{t\in[0,1]}$ is an isotopy from the identity to $F$ on a surface $M$ which has no contractible point, there exists a continuous dynamics on $M$ which is "transverse" to the dynamics of $F$ in the following sense: every orbit is "pushed on its left" by the isotopy. Such a result may be applied in the presence of contractible fixed points if one takes out some of them. Suppose that $F$ is the time-one map of a Hamiltonian flow on a symplectic surface $M$ associated to $H\colon M \to \mathbb{R}$, then an example of a transverse dynamics is the dynamics of the gradient flow of $H$ if we endow $M$ with a Riemannian structure. We will see that such a transverse gradient-like dynamics may be produced even in the time dependent case.

We will recall the classical Brouwer theory of homeomorphisms of the plane in Section 2 and then state equivariant versions in the case of the annulus in Section 3. We will mention some recent results on dynamics of diffeomorphisms and homeomorphisms of surfaces. In particular, in Section 6 we will recall some new results due to Polterovich and to Franks and Handel about group actions on the space of area preserving diffeomorphisms of surfaces.

## 2. Brouwer's theory of plane homeomorphisms

The following result is due to Brouwer and recent proofs are given in [12], [16] or [32].

**Theorem 2.1** ([10]). *Let $f$ be an orientation preserving homeomorphism of the euclidean plane $\mathbb{R}^2$. If $f$ has a periodic point $z$ of period $q \geq 2$, then there is a simple closed curve $\Gamma$, disjoint from the fixed point set* $\mathrm{Fix}(f)$*, such that $i(f, \Gamma) = 1$, where*

*the index* $i(f, \Gamma)$ *is the degree of the map*

$$s \mapsto \frac{f(\Gamma(s)) - \Gamma(s)}{\|f(\Gamma(s)) - \Gamma(s)\|},$$

*and* $s \mapsto \Gamma(s)$ *is a parametrization defined on the unit circle* $S^1$.

A homeomorphism $f$ of $\mathbb{R}^2$ is orientation preserving if and only if it is isotopic to the identity; in the case where $f$ is the time-one map of a flow $(f_t)_{t\in\mathbb{R}}$ whose orbits are tangent to a given continuous vector field $\xi$, the result is obvious. Indeed, if $z$ is a periodic point of $f$ of period $q \geq 2$, the orbit of $z$ (for the flow) is a simple closed curve $\Gamma$, invariant by $f$, which is the union of periodic points of period $q$. This clearly implies that $i(f, \Gamma) = 1$.

Theorem 2.1 asserts that any fixed point free and orientation preserving homeomorphism of $\mathbb{R}^2$ is periodic point free. In fact its dynamics has no recurrence at all. More precisely, suppose that $f$ is an orientation preserving homeomorphism of $\mathbb{R}^2$ and that $f$ has a *non-wandering* point $z$ which is not fixed (i.e. every neighborhood of $z$ meets one of its iterate). Let us see why the conclusion of Theorem 2.1 is still satisfied. Choose a *free* topological open disk $V$ containing $z$ (i.e. disjoint from its image by $f$) and write $q \geq 2$ for the smallest positive integer such that $f^q(V) \cap V \neq \emptyset$. One can compose $f$ with a homeomorphism $h$ supported on $V$ to get a map with a periodic point of period $q$, so that Theorem 2.1 can be applied to $f \circ h$. The map $h$ being supported on a free set, $f \circ h$ and $f$ have the same fixed points. Moreover, $h$ being isotopic to the identity among the homeomorphisms supported on $V$, one has $i(f \circ h, \Gamma) = i(f, \Gamma)$ for every simple closed curve $\Gamma \subset \mathbb{R}^2 \setminus \mathrm{Fix}(f)$. The foregoing argument may be generalized to get the useful Franks Lemma:

**Proposition 2.2** ([21]). *Let* $f$ *be an orientation preserving homeomorphism of* $\mathbb{R}^2$. *If there is a periodic sequence* $(V_i)_{i\in\mathbb{Z}/q\mathbb{Z}}$ *of pairwise disjoint free topological open disks, and a sequence* $(n_i)_{i\in\mathbb{Z}/q\mathbb{Z}}$ *of positive integers such that* $f^{n_i}(V_i) \cap V_{i+1} \neq \emptyset$, *then there is a simple closed curve* $\Gamma \subset \mathbb{R}^2 \setminus \mathrm{Fix}(f)$ *such that* $i(f, \Gamma) = 1$.

It has been known for a long time that Brouwer theory may be applied to the study of homeomorphisms of surfaces. Let us explain for example why every orientation and area preserving homeomorphism on the sphere $S^2$ has at least two fixed points. The map $f$ being orientation preserving has at least one fixed point $z_1$ by the Lefschetz formula. The fact that $f$ preserves the area implies that every point is non-wandering; applying Proposition 2.2 to the map restricted to the topological plane $S^2 \setminus \{z_1\}$, one gets a second fixed point $z_2 \neq z_1$. As noticed by Hamilton [34], then by Brown [11], there is another fixed point result which can be deduced from Proposition 2.2: the Cartwright–Littlewood Fixed Point Theorem [14]. This theorem asserts that any non-separating continuum $K \subset \mathbb{R}^2$ which is invariant by an orientation preserving homeomorphism $f$ of $\mathbb{R}^2$ contains a fixed point. Let us recall Brown's argument. If $K \cap \mathrm{Fix}(f) = \emptyset$, then $K$ is included in a connected component $W$ of $\mathbb{R}^2 \setminus \mathrm{Fix}(f)$. One can choose a lift $\tilde{f}$ of $f|_W$ to the universal covering space $\widetilde{W}$ (homeomorphic

to $\mathbb{R}^2$) which fixes a given connected component $\widetilde{K}$ of the preimage of $K$. The set $\widetilde{K}$ being compact, $\tilde{f}$ should contain a non-wandering point (in fact a recurrent point) while being fixed point free.

Let us state now a much more difficult fixed point theorem, due to Handel, which is very useful in the study of homeomorphisms of surfaces (see [24], [25], [50]):

**Theorem 2.3** ([35]). *Let $f$ be an orientation preserving homeomorphism of the closed unit disk $D$ and suppose that*

- *there are $n \geq 3$ points $z_i$, $1 \leq i \leq n$, in $\mathrm{Int}(D)$ such that $\lim_{k\to-\infty} f^k(z_i) = \alpha_i \in \partial D$ and $\lim_{k\to+\infty} f^k(z_i) = \omega_i \in \partial D$;*
- *the $2n$ points $\alpha_i$ and $\omega_i$ are distinct;*
- *there is an oriented convex compact polygon in $\mathrm{Int}(D)$ whose $i$-th side joins $\alpha_i$ to $\omega_i$.*

*Then $f$ has a fixed point in $\mathrm{Int}(D)$.*

Handel's Fixed Point Theorem is usually applied to a lift $f$ to the universal covering space $\mathrm{Int}(D)$ of a homeomorphism $F$ of a hyperbolic surface (such a lift can always be extended to the closed disk). The core of the proof of Handel is a generalization of the Nielsen–Thurston classification of homeomorphisms of a compact surface $M$ to the case where $M$ is the complement in $\mathbb{R}^2$ of finitely many infinite proper orbits. In fact, it is possible to directly prove Theorem 2.3 by showing the existence of a *periodic free disk chain* of $f|_{\mathrm{Int}(D)}$ (i.e. a family of disks satisfying the assumptions of Proposition 2.2) and thus of a simple closed curve $\Gamma$ of index 1 (see [44]).

We will go on by recalling Brouwer's Plane Translation Theorem. By Schoenflies' Theorem, any proper topological embedding of the real line $\{0\} \times \mathbb{R}$ may be extended to an orientation preserving homeomorphism $h$ of $\mathbb{R}^2$. The open sets $L(\Gamma) = h(]-\infty, 0[\times\mathbb{R})$ and $R(\Gamma) = h(]0, +\infty[\times\mathbb{R})$ are the two connected components of the complement of the *oriented line* $\Gamma = h(\{0\} \times \mathbb{R})$.

**Theorem 2.4** ([10]). *If $f$ is a fixed point free and orientation preserving homeomorphism of $\mathbb{R}^2$, then every point belongs to a Brouwer line, that means an oriented line $\Gamma$ such that $f(\Gamma) \subset L(\Gamma)$ and $f^{-1}(\Gamma) \subset R(\Gamma)$.*

Such a homeomorphism is usually called a *Brouwer homeomorphism*. Observe that $W = \bigcup_{k\in\mathbb{Z}} f^{k+1}(R(\Gamma)) \setminus f^k(R(\Gamma))$ is an invariant open subset homeomorphic to $\mathbb{R}^2$ and that $f|_W$ is conjugate to a non-trivial translation of $\mathbb{R}^2$. Theorem 2.4 asserts that $\mathbb{R}^2$ can be covered by such invariant subsets. The quotient space $\mathbb{R}^2/f$ of orbits of $f$ is a topological surface which is Hausdorff if and only if $f$ is conjugate to a translation. Brouwer homeomorphisms have been studied for a long time. Among the more recent results one may mention the construction of the *oscillating set* by Béguin and Le Roux [3] which is a new topological invariant of Brouwer homeomorphisms. One may also mention the study of *Reeb components* of Brouwer homeomorphisms by Le Roux [48]. Such objects, which generalize the classical Reeb components of

foliations may be defined (in a not easy way) in the framework of Brouwer homeomorphisms and results that are true for foliations may be extended to this discrete case. One may also recall the following results about the topology of the space of Brouwer homeomorphisms when equipped with the compact-open topology: it is arcwise connected and locally contractible [8], more precisely the set of non-trivial affine translations is a strong deformation retract [46].

In the case where $f = f_1$ is the time-one map of a flow $(f_t)_{t\in\mathbb{R}}$ whose orbits are tangent to a continuous vector field $\xi$, Theorem 2.4 is also obvious. Indeed one may find a complete $C^1$ vector field $\eta$ such that $\eta(z) \wedge \xi(z) > 0$ for every $z \in \mathbb{R}^2$ (where $\wedge$ is the usual exterior product on $\mathbb{R}^2$). Every orbit $\Gamma$ of $\eta$ is a line by the Poincaré–Bendixson Theorem; it is a Brouwer line because $\xi$ points on $\Gamma$ from the right to the left. Note that the plane is foliated and not only covered by Brouwer lines. The proof of the Brouwer Plane Translation Theorem is much harder. In all known proofs (see [10], [23], [32], [45]) a non-recurrence lemma, a variation of Proposition 2.2, is needed. We will conclude this section by explaining the ideas of the proofs given in [45] and in Sauzet's thesis [54] as it will be the starting point of the proofs of the foliated versions that we will explain later.

A *brick decomposition* of $\mathbb{R}^2$ is given by a one dimensional stratified set $\Sigma$ (the *skeleton* of the decomposition) with a zero dimensional submanifold $V$ such that any vertex $v \in V$ is locally the extremity of exactly three edges. A *brick* is the closure of a connected component of $\mathbb{R}^2 \setminus \Sigma$. If $f$ is a fixed point free and orientation preserving homeomorphism of $\mathbb{R}^2$, one can construct a *maximal free decomposition*: it is a brick decomposition with free bricks such that the union of two adjacent bricks is no more free. Moreover if $z \in \mathbb{R}^2$ is a given point, one may suppose that $z \in \Sigma$. Let us write $B$ for the set of bricks. A slightly stronger version of Proposition 2.2, due to Guillou and Le Roux [47] asserts that there is no closed chain of bricks of $B$. This implies that the relation

$$b\mathcal{R}b' \iff f(b) \cap b' \neq \emptyset$$

generates by transitivity an order $\leq$ on $B$. The decomposition being maximal, two adjacent bricks are comparable. In fact, it appears that for every brick $b$, the union of bricks $b' > b$ adjacent to $b$ is non-empty and connected, as is the union of adjacent bricks $b' < b$. This implies that $b_\geq = \bigcup_{b'\geq b} b'$ is a connected closed subset satisfying $f(b_\geq) \subset \mathrm{Int}(b_\geq)$. The fact that we are working with bricks implies that the frontier of $b_\geq$ is a one dimensional manifold; the inclusion $f(b_\geq) \subset \mathrm{Int}(b_\geq)$ implies that every component of this frontier is a Brouwer line. One may cover the skeleton by Brouwer lines because $\Sigma = \bigcup_{b\in B} \partial b_\geq$.

Free cellular decompositions appear already in [32] and are explicitely constructed in a paper of Flucher [20] about a topological version of Conley–Zehnder theorem for $\mathbb{T}^2$. The trick to consider bricks to simplify the proofs was suggested by Guillou. Brick decompositions have been studied in detail in Sauzet's thesis [54] and have been used in some articles ([4], [9], [47]). In [47], Le Roux gives a very precise description of the dynamics of a homeomorphism $F$ of a surface in the neighborhood

of an isolated fixed point such that $i(F, z) \neq 1$. In [9], Bonino states the following result about any orientation reversing homeomorphim $f$ of $\mathbb{R}^2$: if $f$ has no periodic point of period 2, it has no periodic point of period $\geq 2$ and the complement of the fixed point set may be covered by invariant open subsets, where $f$ is conjugate either to the map $(x, y) \mapsto (x + 1, -y)$ or to the map $(x, y) \mapsto \frac{1}{2}(x, -y)$.

## 3. Equivariant versions of Brouwer's theory

Let us recall the classical Poincaré–Birkhoff Theorem which is the starting point of Arnold's conjecture:

**Theorem 3.1** ([6]). *Let $F$ be an area preserving homeomorphism of the annulus $\mathbb{T}^1 \times [0, 1]$ isotopic to the identity and let $f$ be a given lift to the universal covering space $\mathbb{R} \times [0, 1]$. Denote by $p_1 \colon \mathbb{R} \times [0, 1] \to \mathbb{R}$ the first projection and suppose that $p_1(f(x, 0)) < x < p_1(f(x, 1))$ for every $x \in \mathbb{R}$. Then $f$ has at least two fixed points which project in different points of $\mathbb{T}^1 \times [0, 1]$.*

Soon after the original proof of Birkhoff, it was noticed (see Birkhoff [7], Kerékjártó [37]) that the existence of one fixed point could be deduced by replacing the area preserving assumption with the following *intersection property*: any essential (i.e. not null-homotopic) simple closed curve of $\mathbb{T}^1 \times [0, 1]$ meets its image by $F$. Kerékjártó obtained the result as a consequence of Brouwer's Plane Translation Theorem. Suppose that $F$ and a given lift $f$ satisfy the assumptions of Theorem 3.1 but the area condition, and that $f$ is fixed point free. Then one can extend nicely $f$ to the whole plane in such a way that a Brouwer line may be constructed on $\mathbb{R}\times]0, 1[$ that is a lift of a simple closed curve of $\mathbb{T}^1\times]0, 1[$ (see [32] for a modern explanation). More recently, Guillou [33] and Sauzet [54] gave a proof of the following "equivariant" version of the Brouwer Plane Translation Theorem:

**Theorem 3.2.** *Let $F$ be a homeomorphism of $\mathbb{T}^1 \times \mathbb{R}$ isotopic to the identity and let $f$ be a given lift to the universal covering space $\mathbb{R}^2$ that is fixed point free. Then*

- *either there is an essential simple closed curve $\Gamma$ of $\mathbb{T}^1 \times \mathbb{R}$ such that $F(\Gamma) \cap \Gamma = \emptyset$,*
- *or there is a topological line joining the two ends of $\mathbb{T}^1 \times \mathbb{R}$ that is lifted to $\mathbb{R}^2$ by Brouwer lines of $f$.*

Franks [21] gave a different method to deduce the Poincaré–Birkhoff Theorem from Brouwer theory by using Proposition 2.2. Suppose that the assumptions of Theorem 3.1 are satisfied and moreover that the number $m$ of fixed points of $F$ which are lifted to fixed points of $f$ is finite. Then one can construct a closed chain of free disks of $f|_{\mathbb{R}\times]0,1[}$, which implies that there is a simple closed curve $\Gamma \subset \mathbb{R}\times]0, 1[$ such that $i(f, \Gamma) = 1$. The Lefschetz–Nielsen formula implies that $m \geq 2$. This argument has been used by Franks to give many generalizations of the Poincaré–Birkhoff Theorem, including results on the torus [22]. We will state below such a

result which is implicitly contained in Franks' papers and which can also be deduced directly from Theorem 3.2 (see [5]).

Under the hypothesis of the Poincaré–Birkhoff Theorem, $F$ admits periodic points of arbitrarily large period. Indeed, if $\rho_-$ and $\rho_+$ are the Poincaré rotation numbers defined respectively on $\mathbb{T}^1 \times \{0\}$ and $\mathbb{T}^1 \times \{1\}$, one has $\rho_- < 0 < \rho_+$. For every rational number $p/q \in ]\rho_-, \rho_+[$ written in an irreducible way, one may apply again the Poincaré–Birkhoff Theorem to $F^q$ and its lift $T^{-p} \circ f^q$ (where $T(x, y) = (x+1, y)$). One gets a fixed point of $T^{-p} \circ f^q$ which projects onto a periodic point of $F$ of period $q$. Moreover, the theory of homeomorphisms of the circle gives us such a periodic point if one of the numbers $\rho_-$ or $\rho_+$ is equal to $p/q$. Let us give now a more general statement. Let us denote by $\mathbb{A}$ one of the annulus $\mathbb{T}^1 \times [0, 1]$ or $\mathbb{T}^1 \times ]0, 1[$ and by $\tilde{\mathbb{A}}$ its universal lift. Write $\pi : \tilde{\mathbb{A}} \to \mathbb{A}$ for the covering projection and $T : (x, y) \mapsto (x+1, y)$ for the fundamental covering automorphism. Consider a homeomorphism $F$ of $\mathbb{A}$ isotopic to the identity and a lift $f$ to $\tilde{\mathbb{A}}$. Suppose that $z \in \mathbb{A}$ is a positively recurrent point of $F$ and that $\tilde{z} \in \tilde{\mathbb{A}}$ is a preimage of $z$. For every sequence $(F^{q_k}(z))_{k\geq 0}$ that converges to $z$, there exists a sequence $(p_k)_{k\geq 0}$ in $\mathbb{Z}$ such that $(T^{-p_k} \circ f^{q_k}(\tilde{z}))_{k\geq 0}$ converges to $\tilde{z}$. The sequence $(p_k)_{k\geq 0}$ is uniquely defined up to a finite number of terms and does not depend on $\tilde{z}$. Let us say that $z$ has a *rotation number* $\rho$ if, for every sequence $(F^{q_k}(z))_{k\geq 0}$ that converges to $z$, the sequence $(p_k/q_k)_{k\geq 0}$ converges to $\rho$. Another choice of lift $f$ changes the rotation number by adding an integer.

**Proposition 3.3.** *Let $F$ be a homeomorphism of $\mathbb{A}$ isotopic to the identity and $f$ a given lift to $\tilde{\mathbb{A}}$. We suppose that*

- *there is a positively recurrent point $z_-$ of rotation number $\rho_-$;*
- *there is a positively recurrent point $z_+$ of rotation number $\rho_+ > \rho_-$;*
- *every essential simple closed curve in $\mathbb{A}$ meets its image by $F$.*

*Then for every rational number $p/q \in ]\rho_-, \rho_+[$ written in an irreducible way, there is a periodic point $z$ of period $q$ and rotation number $p/q$.*

In the case where $F$ is area preserving, the intersection property is satisfied and Proposition 3.3 may be applied. More can be said in that case ([5], [24], [39]). First, one can prove the existence of a periodic point of period $q$ and rotation number $p/q$ if there is a positively recurrent point of rotation number $p/q$. Moreover one can prove the existence of a non-trivial interval of rational rotation numbers if there is a positively recurrent point that has no rotation number. As a consequence, the unique case where such a homeomorphism has no periodic point is the case where there exists an irrational number $\rho$ such that every positively recurrent point (thus almost every point) has a rotation number equal to $\rho$. Such a map is usually called an irrational *pseudo-rotation*.

One may notice the following recent result on irrational pseudo-rotations, previously stated by Kwapisz [38] in the context of the torus, revisited by Béguin, Crovisier,

Le Roux, Patou [4] in the case of a closed annulus and extended by Béguin, Crovisier, Le Roux [5] in the case of an open annulus:

*For any convergent $p/q$ of $\rho$, there exists a simple arc $\gamma$ joining the two ends of the annulus such that the iterates of $\gamma$, $F(\gamma), \dots, F^q(\gamma)$ are pairwise disjoint and cyclically ordered as the iterates of a vertical under a rigid rotation of angle $\rho$.*

As a consequence ([4], [5]) one has the following:

*The rigid rotation of angle $\rho$ is in the closure of the conjugacy class of the pseudo-rotation.*

Note that one does not know if an irrational pseudo-rotation of rotation number $\rho$ is in the closure of the conjugacy class of a rigid rotation of angle $\rho$.

It has been known for a long time that the dynamics of an irrational pseudo-rotation may be strongly different from the dynamics of a rigid rotation. Anosov and Katok [2] gave an example of a smooth ($C^\infty$) irrational pseudo-rotation on the closed annulus which is weakly mixing (and therefore ergodic) relatively to the Lebesgue measure. Many other pathological examples may be constructed as explained by Fayad and Katok in [17]. All these examples are constructed as a limit of diffeomorphisms which are smoothly conjugate to rigid rotations of rational angle. The rotation number is always a Liouville number. In fact, Fayad and Saprykina [18] proved that every Liouville number is the rotation number of a weakly mixing smooth pseudo-rotation on $\mathbb{T}^1 \times [0, 1]$. Such examples do not exist for Diophantine numbers. Indeed, an unpublished result of Herman states that for a smooth diffeomorphism of $\mathbb{T}^1 \times [0, 1]$, the circle $\mathbb{T}^1 \times \{1\}$ is accumulated by a set of positive measure of invariant curves of $F$ if the rotation number induced on $\mathbb{T}^1 \times \{1\}$ is Diophantine. Let us conclude this section by recalling the following old conjecture of Birkhoff, still unsolved, stating that an irrational pseudo-rotation on the closed annulus which is real analytic must be conjugate to a rigid rotation.

## 4. Foliated versions of Brouwer's Plane Translation Theorem

As noticed at the end of Section 2, if $f$ is a Brouwer homeomorphism which is the time-one map of a flow whose orbits are tangent to a continuous vector field $\xi$, then $\mathbb{R}^2$ may be foliated and not only covered by Brouwer lines. Suppose now that $G$ is a discrete group of orientation preserving diffeomorphisms acting freely and properly on $\mathbb{R}^2$ and that $\xi$ is *$G$-invariant* (that means invariant by every element of $G$). By considering the surface $M/G$ one can construct a $G$-invariant $C^1$ vector field $\eta$ satisfying $\eta(z) \wedge \xi(z) > 0$. One deduces that there exists a $G$-invariant foliation by Brouwer lines of $f$. The following result states that this is a general fact:

**Theorem 4.1** ([40], [41]). *If $f$ is a Brouwer homeomorphism, there is an oriented topological foliation $\mathcal{F}$ of $\mathbb{R}^2$ whose leaves are Brouwer lines for $f$. Moreover, if $G$ is a discrete group of orientation preserving homeomorphisms acting freely and properly on $\mathbb{R}^2$ and if $f$ commutes with every $T \in G$, then $\mathcal{F}$ may be chosen $G$-invariant.*

Let us give the idea of the proof of the first statement. Consider the maximal free brick decomposition introduced in Section 2. Using Zorn's Lemma one can extend the order $\leq$ to get a weaker one $\leq'$ which is a total order. If $C = (C_{\leftarrow}, C_{\rightarrow})$ is a cut of $\leq'$ the sets $\bigcup_{b\in C_{\leftarrow}}$ and $\bigcup_{b\in C_{\rightarrow}}$ have the same frontier and the (oriented) frontier of $\bigcup_{b\in C_{\rightarrow}}$ is a union of Brouwer lines because $f\left(\bigcup_{b\in C_{\rightarrow}}\right) \subset \operatorname{Int}\left(\bigcup_{b\in C_{\rightarrow}}\right)$. The set $\mathcal{B}$ of such lines covers the skeleton and may be written $\mathcal{B} = \bigcup_{e\in E} \mathcal{B}_e$, where $E$ denotes the set of edges and $\mathcal{B}_e$ the set of lines $\Gamma \in \mathcal{B}$ containing $e$. One can define a partial order $\preceq$ on the set of oriented lines of $\mathbb{R}^2$: $\Gamma \preceq \Gamma'$ if $R(\Gamma) \subset R(\Gamma')$. The fact that $\leq'$ is a total order implies that two lines of $\mathcal{B}$ do not intersect transversally and consequently that $\preceq$ is a total order when restricted to each $\mathcal{B}_e$. The space $\mathcal{B}$, equipped with the topology generated by the $\mathcal{B}_e$, $e \in E$, is not necessarily Hausdorff but each set $\mathcal{B}_e$ is. In fact each $\mathcal{B}_e$ is compact and the restricted topology coincides with the order topology. As an ordered topological space, $\mathcal{B}$ looks like a *lamination* of $\mathbb{R}^2$, that means a closed subset of leaves of a foliation (in fact it will be isomorphic to a lamination of the foliation that we want to construct). There is a natural (but not unique) way to foliate each brick and then to extend $\mathcal{B}$ by constructing a family of Brouwer lines that cover the plane and that do not intersect transversely. By a desingularization process around each vertex of $\Sigma$, one can blow up our extended family to get a foliation by Brouwer lines.

The proof of the second statement is much harder. First one considers a free brick decomposition invariant by every $T \in G$ and maximal for these properties. It is not necessarily maximal among all the free bricks decomposition; however there is a natural $G$-invariant order $\leq$ on $B$ such that

$$f(b) \cap b' \neq \emptyset \Rightarrow b < b'.$$

Moreover, for every brick $b$, the union of bricks $b' > b$ adjacent to $b$ is non-empty and connected, as is the union of adjacent bricks $b' < b$. As previously one can cover $\Sigma$ by a $G$-invariant family of Brouwer lines. To get our $G$-invariant foliation, one needs to cover $\Sigma$ by a $G$-invariant family of Brouwer lines that do not intersect transversally. If $G$ is abelian (that means if $G = \mathbb{Z}$ or $G = \mathbb{Z}^2$) one knows, by a simple set theory argument, that there is a $G$-invariant total order $\leq'$ weaker than $\leq$: the previous proof is still valid. If $G$ is not abelian, the existence of such an order does not seem so clear. The construction of $\mathcal{B}$ uses more subtle arguments based on the topology of the surface $\mathbb{R}^2/G$.

If $\mathcal{F}$ is an oriented topological foliation of $\mathbb{R}^2$ whose leaves are Brouwer lines of $f$, it is easy to prove that for every point $z$ there is an arc $\gamma\colon [0, 1] \to \mathbb{R}^2$ joining $z$ to $f(z)$ that is *positively transverse* to $\mathcal{F}$. That means that $\gamma$ intersects transversely each leaf that it meets, and locally from the right to the left. One deduces immediately (by lifting the isotopy $(F_t)_{t\in[0,1]}$ to an isotopy $(f_t)_{t\in[0,1]}$ from the identity and applying Theorem 4.1 to $f = f_1$):

**Corollary 4.2.** *Let $(F_t)_{t\in[0,1]}$ be an isotopy from the identity to $F$ on an oriented surface $M$. Suppose that $F$ has no contractible fixed point. Then there exists a*

*topological foliation $\mathcal{F}$ on $M$ that is dynamically transverse to the isotopy: the trajectory $\gamma_z : t \mapsto F_t(z)$ of every point is homotopic, relatively to the extremities, to an arc that is positively transverse to $\mathcal{F}$.*

This result belongs to the category of statements that are obviously true when $F$ is the time-one map of a flow and that can be extended to the case where $F$ is the time-one map of an isotopy from the identity. Let us give now a result that does not. One could ask similarly if there exists a foliation by invariant lines for a Brouwer homeomorphism, as it is true in the case of the time-one map of a flow. The answer is no, there exist Brouwer homeomorphisms without any invariant line (Brown, Slaminka, Transue [13]). Observe that in the case of a flow the foliation by invariant lines was explicitely and uniquely defined. In contrast to this, there are many choices of foliations by Brouwer lines and none of them is canonical.

The only closed surface $M$ where Corollary 4.2 can be applied is the torus $M = \mathbb{T}^2$. Indeed, the Lefschetz–Nielsen formula implies the existence of a contractible fixed point for any homeomorphism isotopic to the identity on an oriented closed surface of genus $\neq 1$. In the case of a torus, a stronger hypothesis on the isotopy will imply additional properties of the foliation. Write $f$ for the natural lift of $F$ to $\mathbb{R}^2$ defined by the isotopy and recall the definition of the *rotation set* $R(f)$ whose origin goes back to Schwartzman [55]. The map $f - \mathrm{Id}_{\mathbb{R}^2}$ is invariant by the integer translations and lifts a continuous function $\psi : \mathbb{T}^2 \to \mathbb{R}^2$. For every Borel probability measure which is invariant by $F$, one may define the *rotation vector* $\rho(\mu) = \int_{\mathbb{T}^2} \psi \, d\mu \in \mathbb{R}^2 \approx H_1(\mathbb{T}^2, \mathbb{R})$ and the set $R(f)$ of rotation vectors of all invariant probability measures. The set $R(f)$ is a non-empty convex compact subset of $H_1(\mathbb{T}^2, \mathbb{R})$. If one supposes that $0 \notin R(f)$ (which of course implies that $f$ is fixed point free) one can find cohomology classes $\kappa \in H^1(\mathbb{T}^2, \mathbb{R})$ that are positive on $R(f)$. One has the following:

**Theorem 4.3** ([43]). *Let $F$ be a homeomorphism of $\mathbb{T}^2$ isotopic to the identity and let $f$ be a lift of $F$ to $\mathbb{R}^2$. Suppose that $\kappa \in H^1(\mathbb{T}^2, \mathbb{R})$ is positive on the rotation set $R(f)$. Then there is a non-vanishing smooth closed 1-form $\omega$ whose cohomology class is $\kappa$, and such that $H(f(z)) - H(z) > 0$ if $H$ is a primitive of the lifted form on $\mathbb{R}^2$.*

The level curves of $H$ define a foliation of $\mathbb{R}^2$ by Brouwer lines of $f$. It projects onto a foliation diffeomorphic to a linear one, the leaves are closed if $\kappa$ is a rational class, they are dense if not. One may ask if a similar statement occurs in higher dimension. Let $(F_t)_{t\in[0,1]}$ be an isotopy from the identity on a compact manifold $M$ and write $\gamma_z : t \to F_t(z)$ for the trajectory of any point $z$. Let $\mu$ be a Borel probability measure invariant by $F$. If $\omega$ is a smooth closed 1-form, the integral $\int_M \left( \int_{\gamma_z} \omega \right) d\mu(z)$ is well defined and vanishes when $\omega$ is exact. As it depends linearly on the cohomology class $[\omega]$ of $\omega$, one may find $\rho(\mu) \in H_1(M, \mathbb{R})$ such that

$$\int_M \left( \int_{\gamma_z} \omega \right) d\mu(z) = \langle [\omega], \rho(\mu) \rangle.$$

The rotation set of the isotopy is the set of rotation vectors $\rho(\mu)$ of invariant probability measures. Here again it is a non-empty convex compact subset of $H_1(M, \mathbb{R})$. Suppose now that $\kappa \in H^1(M, \mathbb{R})$ is positive on the rotation set of the isotopy:

*Does there exist a non-vanishing smooth* 1*-form* $\omega$ *such that* $[\omega] = \kappa$ *and* $\int_{\gamma_z} \omega > 0$ *for every* $z \in M$?

The answer is yes if $F$ is the time-one map of a flow $(F_t)_{t\in\mathbb{R}}$ induced by a smooth vector field $\xi$ (see Fried [30] or Schwartzman [55]). More precisely, $\omega$ may be chosen such that $\langle \omega(z), \xi(z)\rangle > 0$ for every $z \in M$. Fried's proof may be adapted in the discrete case to find a smooth closed 1-form $\omega$ such that $[\omega] = \kappa$ and $\int_{\gamma_z} \omega > 0$ for every $z \in M$. The problem is that $\omega$ can vanish. In the case of a time-one map of a flow, if the rotation set does not contain zero, $\kappa$ may be chosen in $H^1(M, \mathbb{Z})$ and $\omega$ will be written $\omega = dH$ where $H : M \to \mathbb{T}^1$ is a submersion. Consequently $M$ fibers over $\mathbb{T}^1$. Therefore one may naturally ask:

*Suppose that on a given compact manifold* $M$ *one may find an isotopy from the identity whose rotation set does not contain* 0*, does the manifold necessarily fiber over* $\mathbb{T}^1$?

Theorem 4.3 gives us an example where a dynamical assumption on an isotopy implies dynamical properties of some foliation dynamically transverse to the isotopy. In many situations such an assumption will imply dynamical properties of every foliation dynamically transverse. This is the fundamental fact that will permit us to apply Theorem 4.1 and its corollary to the study of homeomorphisms of surfaces. We will conclude this section by an example: a short proof of Proposition 3.3. We will give first two useful statements which illustrate how conservative assumptions satisfied by a homeomorphism can be transposed to dissipative properties of a dynamically transverse foliation. Suppose that $(F_t)_{t\in[0,1]}$ is an isotopy from the identity to $F$ without contractible fixed point on a surface $M$ and that $\mathcal{F}$ is a foliation dynamically transverse. This implies that for every point $z' \in M$ and for every $k \geq 1$ one may find an arc joining $z'$ to $F^k(z')$ that is positively transverse to $\mathcal{F}$. It is easy to prove that this arc may be perturbed into a loop positively transverse to $\mathcal{F}$ if the extremities $z'$ and $F^k(z')$ are sufficiently close to a previously given point $z$. Hence the following holds:

*For every non-wandering point* $z$*, there is a loop based on* $z$ *that is positively transverse to* $\mathcal{F}$.

Fix now a point $z$ and define the set $W$ of points $z' \in M$ which can be joined by an arc from $z$ that is positively transverse to $\mathcal{F}$. It may be noticed that $F(\overline{W}) \subset \mathrm{Int}(W)$. Hence the next assertion is true:

*If every point is non-wandering, then for every points* $z$ *and* $z'$*, there is an arc joining* $z$ *to* $z'$ *that is positively transverse to* $\mathcal{F}$.

Let us now prove Proposition 3.3. Suppose for example that $\rho_- < 0 < \rho_+$ and that $\mathbb{A}$ is open. We want to prove that the intersection property is not satisfied if $f$ has no fixed point. In this case we can construct an oriented foliation $\mathcal{F}$ on $\mathbb{A}$ which is lifted to $\tilde{\mathbb{A}}$ into a foliation by Brouwer lines of $f$. The points $z_-$and $z_+$ being recurrent, there are loops $\Gamma_-$ and $\Gamma_+$ based respectively on $z_-$ and $z_+$ that are

positively transverse to $\mathcal{F}$. Write $[\Gamma] \in H_1(\mathbb{A}, \mathbb{Z})$ for the homology class of any loop $\Gamma$ and consider the generator $[\Gamma_0]$ of the loop $\Gamma_0 : t \mapsto (t + \mathbb{Z}, 1/2)$ defined on $[0, 1]$. The fact that $\rho_- < 0$ implies that $\Gamma_-$ may be chosen such that $[\Gamma_-] = n_-[\Gamma_0]$ where $n_- < 0$. Similarly one may suppose that $[\Gamma_+] = n_+[\Gamma_0]$ where $n_+ > 0$. Using the fact that $\mathcal{F}$ is a non-singular foliation, it is straightforward to prove the following:

- the loops $\Gamma_-$ and $\Gamma_+$ are disjoint;
- there is a unique relatively compact annular component $U$ of $\mathbb{A} \setminus (\Gamma_- \cup \Gamma_+)$;
- the frontier of $U$ is the union of two simple essential loops positively transverse to $\mathcal{F}$;
- the leaves on $\partial U$ are all leaving $U$ or all entering in $U$.

The Poincaré–Bendixson Theorem implies the existence of a closed leaf inside $U$. This leaf does not meet its image by $F$ because it is lifted into a Brouwer line of $f$.

## 5. Hamiltonian homeomorphisms of surfaces

Let us say that a homeomorphism $F$ on an oriented closed surface, time-one map of an isotopy from the identity $(F_t)_{t\in[0,1]}$, is *Hamiltonian* if it preserves a probability measure $\mu$ whose support is $M$ and whose rotation vector is 0. The classical example is obtained when $M$ is endowed with a symplectic structure $\omega$ and when the isotopy is defined by a time dependent Hamiltonian vector field. The measure is nothing but the normalized measure induced by the volume form $\omega$. Let us give another example. Consider an irrational pseudo-rotation $F$ on $\mathbb{T}^1 \times ]0, 1[$ and extend $F$ to the end compactification of the annulus. One gets a Hamiltonian homeomorphism on the sphere that has no periodic points but the two fixed ends. As we will see in this section, extended irrational pseudo-rotations are the only examples, up to conjugacy, of Hamiltonian homeomorphisms having finitely many periodic points.

It was shown by Franks [25] that a Hamiltonian homeomorphism on $S^2$ which has at least three fixed points admits infinitely many periodic points. More recently Franks and Handel [26] proved that a non-trivial Hamiltonian diffeomorphism of a surface of positive genus admits periodic points of arbitrarily large periods (and that this is also the case on a sphere if $F$ has at least three fixed points). Their arguments are mainly of topological nature. The differentiability condition prevents the dynamics to be too wild in a neighborhood of a non-isolated fixed point. For each connected component $U$ of the complement of the fixed point set, they construct a normal form of the restriction map $F|_U$ in the sense of Thurston–Nielsen's theory of homeomorphisms of surfaces like it is usually done for a surface of finite type. There are three cases to look at and in each case periodic orbits may be found for different reasons, the case where there exists at least one pseudo-Anosov component, the case where there is a twist condition in a reducing annulus, the case where the map is

isotopic to the identity. The last case is the most difficult one and subtle geometric arguments that already appeared in [35] are needed.

We will state now a more general result, which gives a positive answer to Conley's conjecture in the case of surfaces:

**Theorem 5.1** ([41], [42]). *Suppose that $F$ is a Hamiltonian time-one map of an isotopy from the identity $(F_t)_{t\in[0,1]}$ on a compact oriented surface $M$ of genus $g \geq 1$.*

i) *If $F \neq \mathrm{Id}_M$, there are periodic points of arbitrarily large period.*

ii) *If the set of contractible fixed points is contained in a disk of $M$, there are contractible periodic points of arbitrarily large period.*

*Moreover we have a similar result in the case where $M$ is a sphere if we suppose that $F$ has at least three fixed points.*

Let us explain first what happens when $F$ is the time-one map of a time independent Hamiltonian flow associated to a function $H : M \to \mathbb{R}$ on a surface of genus $\geq 1$. Let us suppose that there are finitely many critical points of $H$ (there are at least three). The minimum of $H$ corresponds to a contractible fixed point $z_0$. This point is surrounded by invariant curves which are level curves of $H$. The map $F$ is conjugate to a rotation on each curve. Thus one gets a foliated open annulus with one end corresponding to $z_0$ and one "critical" end which does not correspond to a point (because $M$ is not a sphere) but to a degenerate curve containing a critical point of $H$. The rotation number of $F$ on each curve (which is a well defined real number) depends continuously of the curve, never vanishes and tends to zero when the curve tends to the critical level. This implies that the rotation numbers take their values onto a non-trivial interval. One concludes that there are contractible periodic points of arbitrarily large period.

In the case where $F$ is the time-one map of a time dependent Hamiltonian flow, Floer [19] and Sikorav [56] proved that $F$ has at least three contractible fixed points, giving a positive answer to Arnold's conjecture for surfaces. In symplectic geometry contractible fixed points of Hamiltonian isotopies are usually found by studying the dynamics of the gradient flow of a function $\mathcal{H}$ defined on an infinite dimensional space (space of loops) or on a high dimensional space (if one uses generating functions) whose critical points are in bijection with contractible fixed points. Franks [24] gave a purely topological proof of the existence of three contractible fixed points for a Hamiltonian diffeomorphism making use of Handel's Fixed Point Theorem, the proof of which was extended by Matsumoto [50] to the case of Hamiltonian homeomorphisms. The fundamental idea in the proof of Theorem 5.1 is to make a link between the symplectic and the topological methods by producing a "singular" dynamically transverse foliation and by proving that its dynamics is "gradient-like". This will permit us first to find again Matsumoto's result, then to produce a topological "twist property". Such a property is easy to prove if $F$ is a diffeomorphism with no degenerate fixed points. We will give here some ideas of the proof of assertion ii) of Theorem 5.1. We will begin by the simplest case where the set $\mathrm{Fix}(F)_{\mathrm{cont}}$ of contractible fixed point is finite.

*Case where $M = S^2$ and $\sharp\mathrm{Fix}(F)_{\mathrm{cont}} < +\infty$.* Here $\mathrm{Fix}(F)_{\mathrm{cont}}$ coincides with the set $\mathrm{Fix}(F)$ of fixed points. We suppose that $F$ preserves a probability measure $\mu$ with total support and that $3 \leq \sharp\mathrm{Fix}(F) < +\infty$. We want to prove that $F$ has periodic points of arbitrarily large period. Let us say that $Z \subset \mathrm{Fix}(F)$ is *unlinked* if $F$ is isotopic to the identity relatively to $Z$. This is always the case if $\sharp Z \leq 3$. As $\mathrm{Fix}(F)$ is supposed to be finite, one can find a *maximal* (for the inclusion) unlinked set $Z$ and one knows that $\sharp Z \geq 3$. Fix an isotopy $(F_t)_{t\in[0,1]}$ such that $F_t(z) = z$ for every $z \in Z$ and every $t \in [0, 1]$, and look at the restricted isotopy to $N = S^2 \setminus Z$. It is standard to prove that $(F_t|_N)_{t\in[0,1]}$ has no contractible point, by maximality of $Z$. By Corollary 4.2, one may construct a foliation $\mathcal{F}$ on $N$ which is dynamically transverse to the isotopy. As we suppose that $F$ preserves $\mu$ we know that every point is non-wandering, which implies that every point belongs to a loop that is positively transverse to $\mathcal{F}$. This clearly implies that $\mathcal{F}$ has no closed leaf and more generally has only wandering leaves. In fact the dynamics of $\mathcal{F}$ is easy to understand:

- *any leaf $\lambda$ joins a point $\alpha(\lambda) \in Z$ to a different point $\omega(\lambda) \in Z$;*
- *there is no sequence of leaves $(\lambda_i)_{i\in\mathbb{Z}/p\mathbb{Z}}$ such that $\omega(\lambda_i) = \alpha(\lambda_{i+1})$ for any $i \in \mathbb{Z}/p\mathbb{Z}$.*

Fix a leaf $\lambda$ and consider the annulus $A = S^2 \setminus (\alpha(\lambda) \cup \omega(\lambda))$. The isotopy $(F_t|_A)_{t\in[0,1]}$ may be lifted to the universal covering space $\tilde{A}$ of $A$ into an isotopy $(f_t)_{t\in[0,1]}$ from the identity. We will apply Proposition 3.3 by finding two positively recurrent points with different rotation numbers. The map $f = f_1$ clearly fixes every point of the preimage of $Z \setminus (\alpha(\lambda) \cup \omega(\lambda))$, which implies that the rotation vector of any point of $Z \setminus (\alpha(\lambda) \cup \omega(\lambda))$ is 0. The foliation $\mathcal{F}$ is lifted to a foliation on the preimage $\widetilde{N}$ of $N$ which is dynamically transverse to the isotopy $(f_t|_{\widetilde{N}})_{t\in[0,1]}$. Any lift of $\lambda$ is a Brouwer line of $f$ because $\lambda$ joins the two ends of the annulus. It is not difficult, using classical arguments of Ergodic Theory (and in particular the Birkhoff Ergodic Theorem), to prove that $F$ has positively recurrent points whose rotation number is $\neq 0$ (this is the case for almost every point that has a preimage between a given lift $\tilde{\lambda}$ and its image by $f$).

*Case where $g \geq 1$ and $\sharp\mathrm{Fix}(F)_{\mathrm{cont}} < +\infty$.* Here again suppose that the set $\mathrm{Fix}_{\mathrm{cont}}(F)$ of contractible fixed points is finite and say that $Z \subset \mathrm{Fix}_{\mathrm{cont}}(F)$ is unlinked if there is an isotopy $(F_t)_{t\in[0,1]}$ (homotopic to the one given by hypothesis) such that $F_t(z) = z$ for every $z \in Z$ and every $t \in [0, 1]$. Fix a maximal unlinked set $Z$. Again, there exists a foliation $\mathcal{F}$ on $N = M \setminus Z$ which is dynamically transverse to the isotopy $(F_t|_N)_{t\in[0,1]}$, and we would like to understand the dynamics of $\mathcal{F}$. As we suppose that $F$ preserves $\mu$ we already know that every point belongs to a loop positively transverse to $\mathcal{F}$. The fact that the rotation of $\mu$ is zero implies a stronger result:

*For every $\upsilon \in H_1(M, \mathbb{Z})$ and every $z \in M$ there is a loop $\Gamma \subset N$ positively transverse to $\mathcal{F}$ and based in $z$ such that $[\Gamma] = \upsilon$.*

One must prove that the set $C(z) \subset H_1(M, \mathbb{Z})$ of homology classes of loops in $N$ based in $z$ and positively transverse to $\mathcal{F}$, which is stable by addition, is the whole

group $H_1(M,\mathbb{Z})$. The nullity of the rotation vector of $\mu$ implies that every class $\kappa \in H^1(M,\mathbb{R})$ takes different signs on $C(z)$ and therefore that the convex hull in $H_1(M,\mathbb{R})$ of $C(z)$ contains a neighborhood of 0. It becomes easy to prove that $C(z)$ is a subgroup and therefore a lattice of $H_1(M,\mathbb{Z})$. If one now applies the transverse transitivity condition stated in the previous section to a natural finite covering of $M$, one obtains that $C(z) = H_1(M,\mathbb{Z})$.

It is easy to deduce that there is no closed leaf and more precisely that every leaf is wandering. In fact one can prove that the dynamics of $\mathcal{F}$ is gradient-like. Note first that any loop $\Gamma \subset N$ homologous to zero induces naturally by duality a function $\Lambda_\Gamma : M\setminus\Gamma \to \mathbb{Z}$ defined up to a constant, where $\Lambda_\Gamma(z') - \Lambda_\Gamma(z)$ denotes the algebraic intersection number $\Gamma \wedge \Gamma'$ between $\Gamma$ and any arc $\Gamma'$ joining $z$ to $z'$. Observe now that $\Lambda_\Gamma$ decreases along the oriented leaves if $\Gamma$ is positively transverse to $\mathcal{F}$. In other words, the sub-level surfaces of $\Lambda_\Gamma$ define a filtration of $\mathcal{F}$. The property stated above permits us to construct a loop $\Gamma$ homologous to zero and positively transverse to $\mathcal{F}$ which sufficiently "fills" the surface in the following sense:

- *every connected component $U$ of $M \setminus \Gamma$ is the interior of a closed disk of $M$ and contains at most one point of $Z$;*
- *if there exists a leaf of $\mathcal{F}$ which joins $z \in Z$ to $z' \in Z$, then $\Lambda_\Gamma(z') < \Lambda_\Gamma(z)$.*

Using the Poincaré–Bendixson Theorem, one may deduce first that every leaf meets $\Gamma$ and then that it joins a point $z \in Z$ to a point $z' \in Z$. In fact, the dynamics of $F$ is trivial inside a component $U$ with no singularity and well understood inside a component that contains a singularity. Such a singularity is necessarily a sink, a source or a generalized saddle point (with $p \geq 1$ attracting sectors alternating with $p \geq 1$ repelling sectors).

An easy consequence of the previous results is the fact that $\sharp Z \geq 3$. Existence of contractible periodic points of arbitrarily large period is much more difficult to get. One wants to generalize the case where $F$ is the time-one map of a Hamiltonian flow associated to a function $H : M \to \mathbb{R}$. If $M$ is equipped with a Riemannian metric, the foliation by orbits of the gradient flow of $H$ on the complement of the set $Z$ of critical points is dynamically transverse to the isotopy and the point $z_0$ where $H$ reaches its minimum is a sink of the foliation. In our more general situation, one will choose a sink of $\mathcal{F}$ and then will prove that there exists periodic points inside the basin of attraction $W$ (for the foliation). The set $W$ has no reason to be invariant by $F$. However the two following facts

- there exists at least one contractible fixed point in the frontier of $W$,
- there is a radial foliation on $W$ which is pushed along the isotopy,

give us a weak twist condition. Some plane topology arguments and the use of the discrete Conley index permit us to find periodic points inside $W$.

*Case where* $\sharp\mathrm{Fix}(F)_{\mathrm{cont}} = +\infty$. The case where the set of contractible fixed points is infinite is much harder to deal with because it does not seem so easy to find maximal

unlinked sets, which are necessary to construct a dynamically transverse foliation. Under the hypothesis ii) of Theorem 5.1, there is a unique component $N$ of $M \setminus \mathrm{Fix}_{\mathrm{cont}}(F)$ such that the inclusion $i\colon N \to M$ induces an isomorphism between the first groups of homology and this component is fixed. If there is a lift $f$ of $F|_N$ to the universal covering space of $N$ which commutes with every covering transformation, then by Corollary 4.2 a dynamically transverse foliation $\mathcal{F}$ may be constructed. Of course there is no decomposition of the dynamics of $\mathcal{F}$ in elementary pieces as in the finite case. However, the previous arguments may be generalized, even if they are not so easy to get. In the case where such a lift does not exist, we will get contractible periodic orbits of arbitrarily large periods for different reasons, that will be explained in the next section.

There are natural reasons to study carefully homeomorphisms of surfaces of infinite type. Consider a volume form on $S^2$ and write $\mathrm{Diff}^k_\omega(S^2)$ for the set of $C^k$ diffeomorphisms that preserve $\omega$. Consider $F \in \mathrm{Diff}^k_\omega(S^2)$ and fix a connected component $U$ of $S \setminus \overline{\mathrm{Per}(F)}$. There is an integer $q$ such that $F^q(U) = U$. By Theorem 2.1, one knows that there would be a fixed point of $F^q$ in $U$ if $U$ were a disk, which is not the case. By Franks result stated above [25] it cannot be a hyperbolic surface of finite type. Therefore it is an annulus (and in that case the restricted map is an irrational pseudo-rotation) or a surface of infinite type. One may ask the following:

*Can $U$ be a surface of infinite type or should it be necessarily an annulus?*

The interest in this question comes from the following: it is not difficult to prove that there is a residual set $\mathcal{G} \subset \mathrm{Diff}^k_\omega(S^2)$ (for the $C^k$-topology) such that for every $F \in \mathcal{G}$ there are no annulus among the connected components of $S^2 \setminus \overline{\mathrm{Per}(F)}$ (see [29]). A positive answer to the previous question would imply that the periodic orbits are generically dense. What is known is that the union of the stable manifolds of the hyperbolic periodic points is dense [29], a result extended by Xia [57] to any compact surface.

## 6. On the group of diffeomorphisms of surfaces

Consider a compact Riemannian manifold $M$. If $F$ is a $C^1$ diffeomorphism one can define its *growth sequence* $(\Gamma_n(F))_{n\geq 0}$ where

$$\Gamma_n(F) = \max\Big(\max_{z\in M} \|T_z F^n\|,\ \max_{z\in M} \|T_z F^{-n}\|\Big).$$

The growth sequence of a non-trivial diffeomorphism may be bounded. This is the case for a periodic map, a translation on a torus or a rigid rotation on $S^2$. Even when it is not bounded it may tend to $+\infty$ not very quickly (see Polterovitch, Sodin [52]). The situation is different in the case of area preserving diffeomorphisms of surfaces. More precisely:

**Theorem 6.1.** *If $F$ is a non-trivial Hamiltonian diffeomorphism of a closed oriented surface of genus $\geq 1$, there exists $C > 0$ such that $\Gamma_n(F) \geq Cn$ for every $n \geq 0$.*

Proved by Sikorav and Polterovitch in the special case of the torus, the result was generalized to other surfaces by Polterovich [51] applying a result of Schwarz related to Floer homology. Note that in the case of a surface of genus $\geq 2$, the result is still true for any area preserving diffeomorphism isotopic to the identity. Indeed, by the Lefschetz–Nielsen formula, such a map has at least one contractible fixed point. Therefore the diffeomorphism admits two probability measures with different rotation vectors if it is not Hamiltonian. It is not difficult to see that such a property implies that the conclusion of Theorem 6.1 is necessarily true.

Applications of the previous result to actions of higher rank lattices in simple Lie groups on compact manifolds were given in [51], yielding a positive answer, in the special case of surfaces, to a more general conjecture of Zimmer:

**Theorem 6.2.** *Fix a volume form $\omega$ on a closed oriented surface $M$ of genus $g \geq 2$. Then any morphism $\psi$ of $\mathrm{SL}(n, \mathbb{Z})$ in the group $\mathrm{Diff}^{\infty}_{\omega}(M)$ of diffeomorphisms which preserves $\omega$ has a finite image if $n \geq 3$.*

Franks and Handel in [27] gave an alternative proof which works in the $C^1$ case and includes the case $g \leq 1$. The smoothness of $F$ is used in a much weaker way, mainly to construct a Thurston–Nielsen normal form on the complement of the fixed point set. The two important properties satisfied by the group $\mathrm{SL}(n, \mathbb{Z})$, $n \geq 3$, and by any normal subgroup of finite order are the following (the first one is due to Margulis):

- it is almost simple (every normal subgroup is finite or has a finite index);
- it contains a subgroup isomorphic to the group of upper triangular integer valued matrices of order 3 with 1 on the diagonal (the integer Heisenberg group).

More precisely, using algebraic properties of the mapping class group, it is sufficient to study the case where $\psi$ takes its values in the subgroup $\mathrm{Diff}^{1}_{\omega,*}(M)$ of diffeomorphisms of $\mathrm{Diff}^{1}_{\omega}(M)$ which are isotopic to the identity. Using the second property, there exist three elements $F$, $G$, $H$ in $\mathrm{Im}(\psi)$ such that $[G, H] = F$, $[F, G] = [F, H] = \mathrm{Id}_M$ and such that $F$ is the image of an element of infinite order. To get the theorem it is sufficient to prove that $F = \mathrm{Id}_M$, because this would imply that $\mathrm{Ker}(\psi)$, being an infinite normal subgroup, has a finite index. Note that $F$ is Hamiltonian because it is a commutator and that $F^{n^2} = [G^n, H^n]$. The fact that $F = \mathrm{Id}_M$ will follow from the next result (and the fact that $F$ has periodic orbits if it is not trivial):

**Theorem 6.3** ([28]). *Suppose that $F$ is a diffeomorphism of a closed surface $M$ of genus $g$ which satisfies the following distorsion property: it belongs to a finitely generated subgroup of diffeomorphisms isotopic to the identity and there are two sequences $n_k$ and $p_k$ with $p_k = o(n_k)$ and $n_k \to +\infty$ such that $F^{n_k}$ can be written as the product of $p_k$ elements chosen in the (finite) set of generators. Then $F$ is isotopic to the identity relatively to the fixed point set and has no periodic points except the fixed points if $g \geq 2$, if $g = 1$ and $\mathrm{Fix}\, F \neq \emptyset$, or if $g = 0$ and $\sharp\, \mathrm{Fix}\, F \geq 3$.*

The proof uses the Thurston–Nielsen normal form on the complement of the fixed point set explained in Section 5. The distorsion property implies that $F$ is isotopic to the identity relatively to the fixed point set. Every iterate $F^k$ will also satisfy the distorsion property and should be isotopic to the identity relatively to its fixed point set. But this situation cannot occur in case $\mathrm{Fix}(F^k) \neq \mathrm{Fix}(F)$ if $g \geq 2$, if $g = 1$ and $\mathrm{Fix}\, F \neq \emptyset$, or if $g = 0$ and $\sharp\, \mathrm{Fix}\, F \geq 3$.

Let us conclude this article by explaining how to get another interpretation of Theorem 6.1 and Theorem 6.3 with the use of the foliated version of Brouwer's Plane Translation Theorem. We will look at the case of a surface of genus $g \geq 1$ by using the notion of *linking number*. The case of the sphere may be studied in a similar way by using an appropriate notion of linking number.

Suppose that $F$ is the time-one map of an isotopy from the identity $(F_t)_{t\in[0,1]}$ on a closed surface $M$ of genus $g \geq 1$ and lift the isotopy to an isotopy from the identity $(f_t)_{t\in[0,1]}$ on the universal covering space $\widetilde{M}$. One may identify the universal lift $\widetilde{M}$ of $M$ with the complex plane if $g = 1$ or with the Poincaré disk if $g \geq 2$. If $z$ and $z'$ are two fixed points of $f$, the degree of the map $\xi : S^1 \to S^1$ defined by

$$\xi(e^{2i\pi t}) = \frac{f_t(z) - f_t(z')}{|f_t(z) - f_t(z')|}$$

is called the linking number $I(z, z')$ of $z$ and $z'$. One course $I(z, z') = 0$ if $z'$ is the image of $z$ by a covering automorphism. There exits a "natural lift" of $f|_{\widetilde{M}\setminus\{z\}}$ to the universal covering space of the annulus $\widetilde{M} \setminus \{z\}$ which fixes the preimages of every image of $z$ by a covering automorphism of $\widetilde{M}$. The linking number $I(z, z')$ is nothing but the rotation number (up to the sign) of the fixed point $z'$ of $f|_{\widetilde{M}\setminus\{z\}}$ for this natural lift. Note that for every integer $n \geq 1$, the linking number of $z$ and $z'$ for $f^n$ is equal to $nI(z, z')$. In the case where $I(z, z') \neq 0$ it is not difficult to deduce that there exists $C > 0$ such that $\Gamma_n(F) \geq Cn$ for every $n \geq 0$ and also that $F$ does not satisfy the distorsion property in the group of diffeomorphisms isotopic to the identity. Observe that if $F$ preserves a probability measure with total support, then $f|_{\widetilde{M}\setminus\{z\}}$ satisfies the intersection property. Therefore, in this case, if $f$ has two fixed points $z$ and $z'$ such that $I(z, z') \neq 0$, it has periodic points with arbitrarily large period which project onto contractible periodic points of $F$. The next statement permits us to understand why, in the proof of Theorem 5.1, it is sufficient to study the case where the map $F_{|N}$ has a lift to the universal covering space that commutes with the covering transformations.

**Proposition 6.4** ([42]). *Let $(F_t)_{t\in[0,1]}$ be an isotopy from the identity to $F$ on a closed surface $M$ of genus $g \geq 1$ and $(f_t)_{t\in[0,1]}$ the lifted isotopy to the universal covering space $\widetilde{M}$ starting from the identity. Suppose that there is a connected component $N$ of $M \setminus \mathrm{Fix}_{\mathrm{cont}}(F)$ such that the inclusion $i : N \to M$ induces an isomorphism between the first groups of homology and that there is no lift of $F|_N$ to the universal covering space of $N$ that commutes with the covering automorphisms. Then there are two fixed points $z$ and $z'$ of $f = f_1$ such that $I(z, z') \neq 0$.*

Let us give the ideas of the proof. By an approximation argument it is sufficient to study the case where $\mathrm{Fix}_{\mathrm{cont}}(F)$ is finite. One considers a maximal unlinked set $Z \subset \mathrm{Fix}_{\mathrm{cont}}(F)$. By hypothesis one knows that $Z \neq \mathrm{Fix}_{\mathrm{cont}}(F)$. One may suppose that our isotopy $(F_t)_{t\in[0,1]}$ fixes every point of $Z$. We consider a foliation $\mathcal{F}$ on $N' = M \setminus Z$ dynamically transverse to $(F_t|_{N'})_{t\in[0,1]}$ and lift it to a foliation $\widetilde{\mathcal{F}}$ onto the preimage $\widetilde{N}'$ of $N'$ in $\widetilde{M}$. Fix a point $z' \in \mathrm{Fix}(f) \cap \widetilde{N}'$. There is a loop $\Gamma_0 \subset \widetilde{N}'$ based in $z'$ that is positively transverse to $\widetilde{\mathcal{F}}$ and homotopic in $\widetilde{N}'$ to the trajectory $\Gamma_1$ of $z'$. The dual function $\Lambda_{\Gamma_0} : \widetilde{M} \setminus \Gamma_0 \to \mathbb{Z}$ assigning to $z$ the index of $\Gamma_0$ relatively to $z$ is zero outside a compact set and takes finitely many values. One may suppose for example that the maximum $l^+$ of $\Lambda_{\Gamma_0}$ is different from zero. The loop $\Gamma_0$ being positively transverse to the foliation, it is easy to prove that every component of $\widetilde{M} \setminus \Gamma_0$ where $\Lambda_{\Gamma_0}$ takes the value $l^+$ is the interior of a closed disk whose boundary is a simple loop transverse to the foliation. Therefore, there exists a singularity $z$ inside this component. The loop $\Gamma_1$ being homotopic to $\Gamma_0$ in $\widetilde{N}$, the index of $\Gamma_1$ relatively to $z$ is equal to $l^+$. This number is nothing but the linking number $I(z, z')$.

The linking number $I(z, z')$ between a fixed point $z$ and a periodic point $z'$ of $f$ may be defined similarly. The previous proof may be adapted to get:

**Proposition 6.5** ([42]). *Let $(F_t)_{t\in[0,1]}$ be an isotopy from the identity to $F$ on a closed surface $M$ of genus $g \geq 1$ and $(f_t)_{t\in[0,1]}$ the lifted isotopy to the universal covering space $\widetilde{M}$ starting from the identity. For every periodic point $z'$ of $f = f_1$ that is not fixed, there is a fixed point $z$ of $f$ such that $I(z, z') \neq 0$.*

Let us explain how to deduce Theorem 6.1 from Proposition 6.5. Suppose that $F$ is a non-trivial Hamiltonian diffeomorphism of a closed surface $M$ of genus $g \geq 1$. One can choose a periodic point $z'$ of period $\geq 2$. If $z'$ is contractible, then one gets the conclusion of Theorem 6.1 by applying Proposition 6.5. If the rotation vector of $z'$ is non-zero, the conclusion follows easily. It will follow also in the missing case where $z'$ is not contractible but has a rotation vector equal to zero. This is possible only if $g \geq 2$. Identify $\widetilde{M}$ with the Poincaré disk. If $\tilde{z}'$ is a preimage of $z'$ in $\widetilde{M}$, it is not difficult to prove that there exits $C > 0$ such that for every $n \geq 0$ the hyperbolic distance between $z'$ and $f^n(z')$ is minimized by $Cn$, which implies the validity of the conclusion of Theorem 6.1.

The previous arguments imply that a diffeomorphism which satisfies the assumptions of Theorem 6.3 has no periodic points of period $\geq 2$ if $g \geq 2$ or if $g = 1$ and $F$ has a contractible fixed point, and that every fixed point is contractible. One may adapt the arguments of Proposition 6.4 to prove that for every connected component $U$ of $M \setminus \mathrm{Fix}_{\mathrm{cont}}(F)$, the map $F|_U$ has necessarily a lift to the universal covering space which commutes with the covering automorphisms.

The analogs of Theorem 6.2 and Theorem 6.3 for homeomorphisms are unknown. They should be deduced from the (positive) answer to the following open question:

*Suppose that $f$ is the lift to the universal covering space of a homeomorphism $F$ isotopic to the identity defined by an isotopy $(F_t)_{t\in[0,1]}$ and that $f$ has two fixed*

*point $z$ and $z'$ such that $I(z,z') \neq 0$. Does this imply that $F$ does not satisfy the distorsion property in the group of homeomorphisms isotopic to the identity*?

## References


[1] Arnold, V. I., Sur une propriété topologique des applications globalement canoniques de la mécanique classique. *C. R. Acad. Sci. Paris Sér. I Math.* **261** (1965), 3719–3722.

[2] Anosov, D. V., Katok, A. B., New examples in smooth ergodic theory. Ergodic diffeomorphisms. *Trans. Moscow Math. Soc.* **23** (1970), 1–35.

[3] Béguin, F., Le Roux, F., Ensemble oscillant d'un homéomorphisme de Brouwer, homéomorphismes de Reeb. *Bull. Soc. Math. France* **131** (2003), 149–210.

[4] Béguin, F., Crovisier, S., Le Roux, F., Patou, A., Pseudo-rotations of the closed annulus: variation on a theorem of J. Kwapisz. *Nonlinearity* **17** (2004), 1427–1453.

[5] Béguin, F., Crovisier, S., Le Roux, F., Pseudo-rotations of the open annulus. *Bull. Braz. Math. Soc.*, to appear.

[6] Birkhoff, G. D., Proof of Poincaré's last geometric theorem. *Trans. Amer. Math. Soc.* **14** (1913), 14–22.

[7] Birkhoff, G. D., An extension of Poincaré's last geometric theorem. *Acta. Math.* **47** (1925), 297–311.

[8] Bonino, M., Propriétés locales de l'espace des homéomorphismes de Brouwer. *Ergodic Theory Dynam. Systems* **19** (1999), 1405–1423.

[9] Bonino, M., A Brouwer-like theorem for orientation reversing homeomorphisms of the sphere. *Fund. Math.* **182** (2004), 1–40.

[10] Brouwer, L. E. J., Beweis des ebenen Translationssatzes. *Math. Ann.* **72** (1912), 37–54.

[11] Brown, M., A short proof of the Cartwright-Littlewood theorem. *Proc. Amer. Math. Soc.* **65** (1977), 372.

[12] Brown, M., A new proof of Brouwer's lemma on translation arcs. *Houston J. Math.* **10** (1984), 35–41.

[13] Brown, M., Slaminka, E., Transue, W., An orientation preserving fixed point free homeomorphism of the plane which admits no closed invariant line. *Topology Appl.* **29** (1988), 213–217.

[14] Cartwright, M. L., Littlewood, J. C., Some fixed point theorems. *Ann. of Math.* **54** (1951), 1–37.

[15] Conley, C., Zehnder, E., The Birkhoff-Lewis fixed point theorem and a conjecture of V. I. Arnold. *Invent. Math.* **73** (1983), 33–49.

[16] Fathi, A., An orbit closing proof of Brouwer's lemma on translation arcs. *Enseign. Math.* **33** (1987), 315–322.

[17] Fayad, B., Katok, A., Construction in elliptic dynamics. *Ergodic Theory Dynam. Systems* **24** (2004), 1477–1520.

[18] Fayad, B., Saprykina, M., Weak mixing disc and annulus diffeomorphisms with arbitrary Liouville rotation number on the boundary. *Ann. Sci. École Norm. Sup.* (4) **38** (2005), 339–364.

[19] Floer, A., Proof of the Arnold conjectures for surfaces and generalizations to certain Kähler manifolds. *Duke Math. J.* **51** (1986), 1–32.

[20] Flucher, M., Fixed points of measure preserving torus homeomorphisms. *Manuscripta Math.* **68** (1990), 271–293.

[21] Franks, J., Generalizations of the Poincaré-Birkhoff theorem. *Ann. of Math.* **128** (1988), 139–151.

[22] Franks, J., Recurrence and fixed points of surface homeomorphisms. *Ergodic Theory Dynam. Systems* **8*** (1988), 99–107.

[23] Franks, J., A new proof of the Brouwer plane translation theorem. *Ergodic Theory Dynam. Systems* **12** (1992), 217–226.

[24] Franks, J., Rotation vectors and fixed points of area preserving surface diffeomorphisms. *Trans. Amer. Math. Soc.* **348** (1996), 2637–2662.

[25] Franks, J., Area preserving homeomorphisms of open surfaces of genus zero. *New York J. Math.* **2** (1996), 1–19.

[26] Franks, J., Handel, M., Periodic points of Hamiltonian surface diffeomorphisms. *Geom. Topol.* **7** (2003), 713–756.

[27] Franks, J., Handel, M., Area preserving group actions on surfaces. *Geom. Topol.* **7** (2003), 757–771.

[28] Franks, J., Handel, M., Distortion elements in group actions on surfaces. *Duke Math. J.* **131** (2006), 441–468.

[29] Franks, J., Le Calvez, P., Regions of instability for non-twist maps. *Ergodic Theory Dynam. Systems* **23** (2003), 111–141.

[30] Fried, D., The geometry of cross sections to flows. *Topology* **21** (1982), 353–371.

[31] Fukaya, K., Ono, K., Arnold conjecture and Gromov-Witten invariant. *Topology* **38** (1999), 933–1048.

[32] Guillou, L., Théorème de translation plane de Brouwer et généralisations du théorème de Poincaré-Birkhoff. *Topology* **33** (1994), 331–351.

[33] Guillou, L., Free lines for homeomorphisms of the open annulus. Preprint.

[34] Hamilton, O. H., A short proof of the Cartwright-Littlewood fixed point theorem. *Canad. J. Math.* **6** (1954), 522–524.

[35] Handel, M., A fixed point theorem for planar homeomorphisms. *Topology* **38** (1999), 235–264.

[36] Hofer, H., Zehnder, E., *Symplectic invariants and Hamiltonian dynamics*. Birkhäuser Adv. Texts, Birkhäuser, Basel 1994.

[37] Kerékjártó, B., The plane translation theorem of Brouwer and the last geometric theorem of Poincaré. *Acta Sci. Math. Szeged* **4** (1928-1929), 86–102.

[38] Kwapisz, J., Combinatorics of torus diffeomorphisms. *Ergodic Theory Dynam. Systems* **23** (2003), 559–586.

[39] Le Calvez, P., Rotation numbers in the infinite annulus. *Proc. Amer. Math. Soc.* **129** (2001), 3221–3230.

[40] Le Calvez, P., Une version feuilletée du théorème de translation de Brouwer. *Comment. Math. Helv.* **79** (2004), 229–259.

[41] Le Calvez, P., Une version feuilletée équivariante du théorème de translation de Brouwer. *Inst. Hautes Études Sci. Publ. Math.* **102** (2006), 1–98.

[42] Le Calvez, P., Hamiltonian homeomorphisms of surfaces. *Duke Math. J.*, to appear.

[43] Le Calvez, P., Multivalued Lyapounov functions for homeomorphisms of the 2-torus. *Fund. Math.* **189** (2006), 227–253.

[44] Le Calvez, P., Une démonstration directe du théorème de point fixe de Handel. Preprint.

[45] Le Calvez, P., Sauzet, A., Une démonstration dynamique du théorème de translation de Brouwer. *Exposition. Math.* **14** (1996), 277–287.

[46] Le Roux, F., Étude topologique de l'espace des homéomorphismes de Brouwer (I), (II). *Topology* **40** (2001), 1051–1087; ibid. 1089–1121.

[47] Le Roux, F., Homéomorphismes de surfaces, théorème de la fleur de Leau-Fatou et de la variété stable. *Astérisque* **292** (2004).

[48] Le Roux, F., Structure des homéomorphismes de Brouwer. *Geom. Topol.* **9** (2005), 1689–1774.

[49] Liu, G., Tian, G., Floer homology and Arnold conjecture. *J. Differential Geom.* **49** (1998), 1–74.

[50] Matsumoto, S., Arnold conjecture for surface homeomorphisms. In *Proceedings of the French–Japanese Conference "Hyperspace Topologies and Applications"* (La Bussière, 1997), *Topology. Appl.* **104** (2000), 191–214.

[51] Polterovich, L., Growth of maps, distortion in groups and symplectic geometry. *Invent. Math.* **150** (2002), 655–686.

[52] Polterovich, L., Sodin, M., A growth gap for diffeomorphisms of the interval. *J. Anal. Math.* **92** (2004), 191–209.

[53] Salamon, D., Zehnder, E., Morse theory for periodic solutions of Hamiltonian systems and the Maslov index. *Comm. Pure Appl. Math.* **45** (1992), 1303–1360.

[54] Sauzet, A., Application des décompositions libres à l'étude des homéomorphismes de surface. Thèse de l'Université Paris 13, 2001.

[55] Schwartzman, S., Asymptotic cycles. *Ann. of Math.* **68** (1957), 270–284.

[56] Sikorav, J.-C., Points fixes d'une application symplectique homologue à l'identité. *J. Differential Geom.* **22** (1985), 49–79.

[57] Xia, Z., Area-preserving surface diffeomorphisms. Preprint.

Laboratoire Analyse Géométrie et Applications, C.N.R.S.-U.M.R 7539, Institut Galilée,
Université Paris 13, 99 Avenue J.-B. Clément, 93430 Villetaneuse, France

E-mail: lecalvez@math.univ-paris13.fr

# All, most, some differentiable dynamical systems

Michael Shub*

**Abstract.** In the first part of this paper we study dynamical systems from the point of view of algebraic topology. What features of all dynamical systems are reflected by their actions on the homology of the phase space? In the second part we study recent progress on the conjecture that most partially hyperbolic dynamical systems which preserve a smooth invariant measure are ergodic, and we survey the known examples. Then we speculate on ways these results may be extended to the statistical study of more general dynamical systems. Finally, in the third part, we study two special classes of dynamical systems, the structurally stable and the affine. In the first case we study the relation of structural stability to entropy, and in the second we study stable ergodicity in the homogeneous space context.

**Mathematics Subject Classification (2000).** Primary 37; Secondary 37C, 37D.

**Keywords.** Dynamical system, entropy, Entropy Conjecture, partial hyperbolicity, accessibility, ergodicity, Lyapunov exponent, SRB measure, structural stability, affine diffeomorphism.

## 1. Introduction

We study discrete differentiable dynamical systems $f: M \to M$ on a smooth closed manifold of dimension $m$.[†] Thus, $f \in \mathrm{Diff}^r(M)$ or $\mathrm{End}^r(M)$, the $C^r$ diffeomorphisms or endomorphisms of $M$ respectively, where $1 \leq r \leq \infty$, and occasionally, $r = 0$.

What can be said about differentiable dynamical systems? The best things that can be said concern *all* systems. When we cannot make statements about all systems we may content ourselves with *most* systems. We expect that properties which hold for most systems hold for a specific system under consideration, but we cannot be sure until we have proven it.

Section 2 concerns properties which may hold for *all* dynamical systems, mainly properties from algebraic topology. Principal among these is the Entropy Conjecture which relates the topological entropy of a dynamical system to the induced map on the homology groups.

In Section 3 we turn from *all* to *most*. We investigate the time honored role of (a) *some* hyperbolicity, especially as it concerns (b) the stable and unstable manifolds

*The author would like to thank Charles Pugh for years of collaboration and also for help in preparing this article.

[†] "Closed" means that $M$ is compact and has empty boundary.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of points, (c) their intersections and (d) the equivalence relation these intersections define in the manifold. In the by now classical uniformly hyperbolic case, the equivalence classes form Smale's spectral decomposition and the behavioral properties entailed are structural stability, SRB measures, and ergodicity in the volume preserving Anosov case.

Uniformly hyperbolic systems are *some*, not *most* dynamical systems. So from the point of view of hoping to describe most dynamical systems we relax the structural properties to *some* hyperbolicity . Our goal is to understand how hypotheses about (a)–(d) affect ergodicity of volume preserving diffeomorphisms and whether these hypotheses hold for most partially hyperbolic volume preserving diffeomorphisms. Later we speculate on how they may affect the existence of SRB measures. Our theme is that a little hyperbolicity goes a long way toward ergodicity. Part of our problem is that the (un)stable manifolds, their intersections, and the equivalences they define are topological objects, while the desired results we wish to conclude are measure theoretic. Working in mixed categories raises rather severe technical difficulties, some of which have only recently been overcome.

We conjecture that most volume preserving partially hyperbolic dynamical systems (initially studied by Brin and Pesin) are ergodic, and we survey the rather substantial recent results in this direction, especially by Keith Burns and Amie Wilkinson, and Federico and Jana Rodriguez Hertz and Raul Ures. Here we first confront the role of the equivalence relation on $M$ induced by the strong stable and unstable manifolds and their intersections. This equivalence relation divides the manifold into *accessibility classes*. The main problem is to understand the relationship of the topologically defined accessibility classes of a partially hyperbolic dynamical system to the measure theoretically defined ergodic components via the Anosov–Hopf argument for ergodicity.

In Section 4.1 we study flows on homogeneous spaces and more generally affine diffeomorphisms. The ergodic theory of affine diffeomorphisms and flows on homogeneous spaces is extremely well developed. It relies to a large extent on the structure of Lie groups and representation theory. The ergodicity results in Section 3 apply outside of the homogeneous space context and per force use different techniques such as the accessibility relationship and *julienne quasi-conformality*. Juliennes are dynamically defined sets and quasi-conformality applies to the holonomy maps of the invariant stable and unstable manifolds. How good are these techniques when applied back in the homogeneous space context where a more elaborate set of tools is available for the study of ergodicity and stable ergodicity? While the proofs are very different there is a remarkable coincidence between those affine diffeomorphisms which are stably ergodic when considered with respect to affine perturbations and those which are stably ergodic with respect to all perturbations. Some rather interesting cases remain unresolved. The coincidence of results makes us feel that we have landed in the right place with our definitions of accessibility and makes the outstanding cases even more interesting.

In Section 4.2 we see how the results of Section 2 and 3 relate to one another. The SRB measures were initially proven to exist for uniformly hyperbolic dynamical systems. The Entropy Conjecture holds for these diffeomorphisms and we consider how sharp it is. How much complexity must a diffeomorphism have beyond that which is forced by the Entropy Inequality? Of particular interest are the Morse–Smale diffeomorphisms. The study of these diffeomorphisms has a deep connection to the theory of the structure of manifolds in high dimensions accomplished by Smale. Yet there are new invariants and obstructions.

The relations between dynamics and algebraic topology studied in Sections 2 and 4.2 *may* hold for all $r \geq 1$ but there are definite distinctions between the ergodic theory of $C^1$ and $C^2$ dynamical systems, so in Sections 3 and 4.1 we mostly assume that $r \geq 2$. Sections 2 and 4.2 and Sections 3 and 4.1 may be read independently of one another. But I think it would be a mistake to disassociate them. For one thing, the hyperbolic systems are partially hyperbolic. To understand the partially hyperbolic we must first understand the hyperbolic. For another, the variational principle ties measure theoretic entropy to topological entropy. (See for example Problem 3 of Section 2.) One of the main themes of this talk are the structures that link and the ties that bond the topological and measure theoretic in the presence of smoothness and some hyperbolicity. Moreover, what is true for *all* must be taken into consideration when studying *most*.

## 2. *All* differentiable dynamical systems

What dynamical properties hold for *all* dynamical systems $f$? The answer often depends on the degree of differentiability of $f$.

- Every continuous dynamical system supports an invariant probability measure.
- Every Lipschitz dynamical system has finite topological entropy, but non-Lipschitz systems can have infinite topological entropy.
- Every $C^\infty$ dynamical system satisfies the Entropy Inequality explained below, but this can fail for Lipschitz dynamical systems that are not continuously differentiable.

Let us recall the concept of entropy and the statement of the Entropy Conjecture. The topological entropy of $f$ measures the growth rate of its epsilon distinguishable orbits. It makes sense for any continuous endomorphism of a compact metric space, $f\colon X \to X$. Given $\epsilon > 0$ and $n \in \mathbb{N}$, let $N(f, n, \epsilon)$ be the maximum cardinality of a subset $A \subset X$ such that for each pair of distinct points $x, y \in A$ there is an iterate $f^j$ with $0 \leq j \leq k$ and

$$d(f^j(x),\, f^j(y)) > \epsilon.$$

Then, $h(f, \epsilon)$ is the exponential growth rate of $h(f, n, \epsilon)$ as $n \to \infty$, namely

$$h(f, \epsilon) = \limsup_{n\to\infty} \frac{1}{n} \ln h(f, n, \epsilon).$$

The supremum of $h(f, \epsilon)$ over all $\epsilon > 0$, or what is the same thing, its limit as $\epsilon \to 0$, is the topological entropy of $f$, $h(f)$. In [New1], Newhouse surveys how the concept of entropy fits into the $C^r$ category.

There is a corresponding growth rate in algebraic topology. The map $f\colon M \to M$ induces a homology homomorphism $f_*\colon H_*(M, \mathbb{R}) \to H_*(M, \mathbb{R})$. Under $f_*^n$, homology classes grow no more rapidly than $s^n$ where $s = s(f_*)$ is the spectral radius of $f_*$, i.e., the modulus of the largest eigenvalue of $f_{*i}\colon H_i(M, \mathbb{R}) \to H_i(M, \mathbb{R})$, $0 \le i \le m$.

**Entropy Conjecture** ([Sh2]). For all $C^r$ dynamical systems, $r \ge 1$, we have the *Entropy Inequality*

$$h(f) \ge \ln s(f_*).$$

Of course, the conjecture for $r = 1$ implies all the others, so this is the principal case. But if it fails for $r = 1$ and holds for larger $r$, this is also interesting. The Entropy Conjecture is true for $C^\infty$ dynamics, but remains unknown for $C^r$ dynamics, $1 \le r < \infty$. The positive result is due to Yomdin, [Yom], who compares the growth rate of the volumes of submanifolds of $M$ under iteration of $f$ to the entropy. See also [Gro2].

The Entropy Conjecture is in general false for Lipschitz endomorphisms already on the 2-sphere, and also for Lipschitz or piecewise linear homeomorphisms in dimension four or larger, [Pu]. For $C^1$ $f$, Misiurewicz and Przytycki [MiPr] prove that $h(f) \ge \ln(\text{degree}(f_{*m}))$. Some entropy lower bounds are known for continuous endomorphisms in terms of the growth rate of the induced map on the fundamental or first homology group, [Ma1], [Bo], [FaSh]. These imply entropy lower bounds for homeomorphisms of manifolds below dimension 4 by Poincaré duality. See [MaPr] for recent results.

Here are some more problems which are of a similar nature, relating algebraic topology to differentiable dynamics. We use the notation

$$\text{GR}(a_n) = \limsup_{n\to\infty} \frac{1}{n} \ln a_n$$

to denote the exponential growth rate of a sequence $(a_n)$ in $(0, \infty]$.

Let $V$ and $W$ be closed smooth submanifolds of complementary dimension in the closed manifold $M$, and let $f$ be a smooth endomorphism of $M$. Let $N_n$ denote the number of distinct points of intersection of $f^n(V)$ with $W$ and let $I_n$ denote the intersection of the homology classes $f_*^n[V]$ and $[W]$, where $[V]$ and $[W]$ are the homology classes in $M$ represented by $V$ and $W$.

**Problem 1.** Is $\text{GR}(N_n) \ge \text{GR}(I_n)$?

A special case of this problem concerns the Lefschetz formula. Let $N_n(f)$ be the number of geometrically distinct periodic points of $f$ of period $n$. Let $L(f^n) = \sum_{i=0}^m (-1)^i \,\text{trace}(f_{*i}\colon H_i(M) \to H_i(M))$.

**Problem 2.** Is $\mathrm{GR}(N_n(f)) \geq \mathrm{GR}(|L(f^n)|)$?

By the transversality theorem the inequalities in the last two problems hold $C^r$ generically. The question is: Do they always hold? It is known that if $f$ is $C^1$ and $L(f^n)$ is unbounded then so is $N_n(f)$ [ShSu1]. This fails for Lipschitz maps.

A first interesting case is a smooth degree two map, $f$, of the 2-sphere. Let $N_n$ be the number of distinct periodic points of $f$ of period $n$.

**Problem 3.** Is $\mathrm{GR}(N_n) \geq \ln 2$?

The results of [MiPr] concerning topological entropy and degree and of Katok [Ka] comparing $\mathrm{GR}(N_n)$ to topological entropy for diffeomorphisms in dimension 2 make a start on this problem.

All these examples fall into the following general framework. Let $F$ be a functor from the category of manifolds to another category. Since a dynamical system $f$ may be iterated so may $F(f)$. We ask to compare the asymptotic behavior of the iterates of $F(f)$ and $f$. Here, we considered the functors of algebraic topology. Later the structures we consider and questions we ask for most or some $f$ consider functors such as the tangent bundle, measures, the de Rham complex, etc.

## 3. *Most* differentiable dynamical systems

Since the range of dynamical behavior exhibited by all dynamical systems seems too large to admit a meaningful universal description applicable to all systems, many attempts have been made to describe features of *most* dynamical systems. SRB measures were introduced by Sinai, Ruelle and Bowen in the 1970s in the study of uniformly hyperbolic dynamical systems. The space integrals for continuous functions with respect to these measures predict the time averages of almost every Lebesgue point in the manifold. It is a fundamental result of Sinai, Ruelle and Bowen [Si], [Ru1], [BoRu] that a finite number of SRB measures exist for $C^2$ hyperbolic dynamics (technically Smale's Axiom A and no cycle systems.) Ruelle [Ru2] suggested that these measures apply much more generally. Much effort in dynamical systems in recent years has focused on Ruelle's suggestion. One widespread optimistic program dating from the late 1970s suggests that most systems have a finite (or perhaps countable) collection of ergodic SRB measures. For volume preserving diffeomorphisms of closed manifolds this program can not be correct because the KAM phenomenon insures the robust existence of positive measure sets of codimension one tori with quasi-periodic motions [ChSu], [Yoc], [Xi]. These tori have no non-zero Lyapunov exponents. So the existence of some non-zero exponents may be decisive for the program.

**3.1. Partially hyperbolic diffeomorphisms.** In contrast, we have suggested that a little hyperbolicity goes a long way towards ergodicity of volume preserving diffeomorphisms and hence (trivially) a unique SRB measure. Concretely our principal

results are limited to $C^2$ partially hyperbolic volume preserving diffeomorphisms. These systems are generalizations of Anosov (globally hyperbolic) dynamical systems. In the Anosov case volume preserving $C^2$ diffeomorphisms are proved to be ergodic [An], [AnSi], [Ho]. Brin and Pesin [BrPe] studied ergodicity of partially hyperbolic diffeomorphism with an *accessibility* property. The hypotheses of their ergodicity theorem were too limiting to be broadly applicable. In fact they probably almost never hold, [ShWi2], [HiPe]. In a series of papers [GrPuSh] , [Wi1], [PuSh3], [PuSh4], [PuSh5], [BuWi2], [BuWi3], [RHRHUr] these hypotheses have been replaced by ones quite generally applicable.

More precisely:

**Definition.** A diffeomorphism $f: M \to M$ is *partially hyperbolic* if there is a continuous $Tf$-invariant splitting $TM = E^{\mathrm{u}} \oplus E^{\mathrm{c}} \oplus E^{\mathrm{s}}$ such that $Tf$ is hyperbolic on $E^{\mathrm{u}} \oplus E^{\mathrm{s}}$ and the hyperbolicity dominates $Tf$ on $E^{\mathrm{c}}$ in the sense that for some $\tau$, $\lambda$ with $1 \leq \tau < \lambda$ and positive constants $c$, $C$ we have the following:

(a) For all $v \in E^{\mathrm{u}}$ and all $n \geq 0$, $c\lambda^n|v| \leq |Tf^n(v)|$.

(b) For all $v \in E^{\mathrm{s}}$ and all $n \geq 0$, $|Tf^n(v)| \leq C\lambda^{-n}|v|$.

(c) For all $v \in E^{\mathrm{c}}$ and all $n \geq 0$, $c\tau^{-n}|v| \leq |Tf^n(v)| \leq C\tau^n|v|$.

(d) The bundles $E^{\mathrm{u}}$, $E^{\mathrm{s}}$ are non-zero.

Condition (d) is present to avoid triviality. Without it, every diffeomorphism would be partially hyperbolic, for we could take $E^{\mathrm{c}}$ as $TM$. Sometimes, one only requires $E^{\mathrm{u}} \oplus E^{\mathrm{s}} \neq 0$, but for simplicity we use the stronger assumption (d) in this paper.

Partial hyperbolicity means that under $Tf^n$, vectors in $E^{\mathrm{c}}$ grow or shrink more gradually than do vectors in $E^{\mathrm{u}}$ and $E^{\mathrm{s}}$. The center vectors behave in a relatively neutral fashion. The definition can be recast in several different ways. For instance, expansion of $E^{\mathrm{u}}$ under positive iteration of $Tf$ can be replaced by contraction under negative iteration. Also, non-symmetric rates can be used for expansion and contraction. More significantly, one could permit pointwise domination instead of the absolute domination as above. See [Puj], [BoDíVi] for a discussion of dominated splitting. All of these refinements to the notion of partial hyperbolicity are exploited by Burns and Wilkinson in their result discussed below.

Given a smooth manifold $M$, fix a smooth volume $\mu$ on $M$. Then we say $f$ is volume preserving if it preserves this volume and we write the set of $\mu$ preserving $C^r$ diffeomorphisms of $M$ as $\mathrm{Diff}^r_\mu(M)$.

A diffeomorphism is *ergodic* if it preserves a measure and each measurable invariant set is a zero set or the complement of a zero set. No measurable invariant set has intermediate measure. Ergodicity is *stable* if it persists under perturbation of the dynamical system. Towards our theme that a little hyperbolicity goes a long way toward ergodicity and more optimistically toward the goal of finding SRB measures, we have our main conjecture.

**Main Conjecture.** Among the volume preserving $C^r$ partially hyperbolic dynamical systems for $r \geq 2$, the stably ergodic ones form an open and dense set.

An approach to the Main Conjecture via two additional conjectures consists in generalizing the Anosov–Hopf proof of the ergodicity of Anosov systems ($E^c = \{0\}$) by studying the accessibility relationship. The Anosov–Hopf argument proceeds as follows. If $x, y$ are forward asymptotic then the time average of continuous functions along the orbit of $x$ equals the time average along the orbit of $y$. Reversing time, the same is true for $f^{-1}$ and points $x, y$ which are asymptotic in negative time. Now the Birkhoff ergodic theorem says that positive time averages equal negative time averages almost everywhere. So we say $x \sim y$ if $x$ and $y$ are positive or negative asymptotic and extend $\sim$ to an equivalence relation on $M$. In principle by the Birkhoff theorem time averages should be constant on equivalence classes and we may prove ergodicity by proving that the equivalence classes are measure zero or one. There are severe technical difficulties to this program but it can be made to work in the Anosov and the partially hyperbolic cases with some extra hypotheses. We say $x, y \in M$ are *us-accessible* if there is a piecewise differentiable path joining $x$ to $y$ and tangent either to $E^u$ or $E^s$ at every point of differentiability. A diffeomophism is *e-*(*ssentially*) *accessible* (in the measure theoretic sense) if the only subsets of $M$ saturated with respect to us-accessibility have measure 0 or 1. A diffeomorphism is *us-accesssible* if $M$ itself is a us-accessibility class. us-accessibility obviously implies e-accessibility.[‡]

**Conjecture A.** Every $C^2$ volume preserving e-accessible partially hyperbolic diffeomorphism is ergodic.

**Conjecture B.** The partially hyperbolic diffeomorphisms with the us-accessibility property are open and dense in the $C^r$ partially hyperbolic diffeomorphisms for every $r \geq 1$, volume preserving or not.

Conjectures A and B obviously imply the main conjecture.

Conjecture A was proven with two technical hypotheses in [PuSh4], center bunching and dynamical coherence. Burns and Wilkinson [BuWi2], [BuWi3] have since removed the dynamical coherence hypothesis and improved the center bunching condition. The center bunching condition puts bounds on the ratios of the expansions and contractions in $E^u$ and $E^s$ as compared to $E^c$. If $Tf|_{E^c}$ is close to conformal the center bunching conditions are satisfied.[§]

[‡]Note that the us-accessibility classes are contained in the $\sim$ equivalence classes we defined above. They are much more amenable to use in proofs.

[§]Burns and Wilkinson's center bunching conditions suppose that there are continuous positive functions $\nu(p), \hat{\nu}(p), \gamma(p), \hat{\gamma}(p)$ such that for every $p \in M$:

1. $\nu(p), \hat{\nu}(p) < 1$ and $\nu(p) < \gamma(p) < \hat{\gamma}(p)^{-1} < \hat{\nu}(p)^{-1}$.
2. $\| T_p f(v) \| < \nu(p)$ for $v \in E^s(p)$,
   $\gamma(p) < \| T_p f(v) \| < \hat{\gamma}(p)^{-1}$ for $v \in E^c(p)$,
   $\| T_p f(v) \| > \hat{\nu}(p)^{-1}$ for $v \in E^u(p)$ .
3. $\nu(p) < \gamma(p)\hat{\gamma}(p)$ and $\hat{\nu}(p) < \gamma(p)\hat{\gamma}(p)$.

The second condition is the partial hyperbolicity and the third the center bunching.

We say that $f$ is BW partially hyperbolic and center bunched, if it satisfies the Burns–Wilkinson conditions.

**Theorem** (Burns–Wilkinson [BuWi3]). *Let $f$ be $C^2$, volume preserving, BW partially hyperbolic and center bunched and essentially accessible. Then $f$ is ergodic and in fact a K-automorphism.*

When the dimension of the center bundle $E^c$ is one the bunching conditions are automatically satisfied. So it follows as a simple corollary that:

**Corollary** (Burns–Wilkinson). *Conjecture A is true when dimension $E^c$ is one.*

Even more is true when the dimension of the center bundle $E^c$ is one, Federico and Jana Rodrigues Hertz, and Raul Ures prove the Main Conjecture.

**Theorem** ([RHRHUr]). *When the dimension of $E^c$ is one, Conjecture A, Conjecture B for volume preserving diffeomorphisms and hence the Main Conjecture all are true.*

Towards Conjecture B in general there is [DoWi] in the $C^1$ topology.

The major new elements in the proofs of the series of theorems on stable ergodicity of partially hyperbolic systems are dynamically defined sets called juliennes which can be used to estimate Lebesgue volumes either directly or by proving that they form a Lebesgue density basis and an analysis of the stable and unstable holonomy maps which are julienne quasi-conformal.

Partial hyperbolicty and center bunching are easily seen to be open conditions and us-accessibility is frequently easily proven to hold in an open neighborhood of a given example. Sometimes even e-accessibility is (not so easily) proved to hold in the neighborhood of a given example [RH]. The situation is good enough to be able to conclude stable ergodicity in the $C^2$ topology of quite a few examples. Here are several examples. See [BuPuShWi], [PuSh5] for more details and for more on the current state of affairs.

1. The product of a volume preserving Anosov diffeomorphism and any other volume preserving diffeomorphism can be arbitrarily $C^\infty$ closely approximated by a partially hyperbolic, us-accessible stably ergodic diffeomorphism [ShWi1], [BuPuShWi], as long as the hyperbolicity of the Anosov diffeomorphism is strong enough to produce a partially hyperbolic splitting of the tangent bundle. (Conjecturally an open and dense set of perturbations is ergodic.) *So the KAM phenomenon seems to be dominated by the hyperbolic phenomenon and ergodicity of weakly coupled systems of KAM and Anosov type should be expected to be ergodic.*

2. The time $t$ map of the geodesic flow of a manifold of negative curvature is stably ergodic.

3. Skew products which are compact group extensions over standard Anosov diffeomorphisms are generically us-accessible and $C^2$ stably ergodic, [Br1], [Br2], [BuWi1], [FiPa].

4. Ergodic toral automorphisms having a two dimensional invariant subspace with isometric derivative and some mild extra technical conditions are $C^r$ stably ergodic for a fairly large $r$ [RH].

5. Partially hyperbolic affine diffeomorphisms of finite volume compact homogeneous spaces of simple Lie groups are stably ergodic. We discuss these below.

Systems whose Lyapunov exponents are non-zero, called non-uniformly hyperbolic, were introduced by Pesin and play a large role in the ergodic theory of volume preserving diffeomorphisms and the study of SRB measures. Pesin's paper [Pe1] raises the question if in dimension bigger than two those diffeomorphisms without zero Lyapunov exponents are generic. We have mentioned above that KAM theory produces open sets of volume preserving diffeomorphisms with positive measure sets of invariant tori which have no hyperbolicity. So the answer to the question is "no". But it may be an either/or situation.

**Problem 4** ([ShWi2]). Is it true for generic $f \in \mathrm{Diff}^r_\mu(M)$ that for almost every ergodic component of $f$ either all the Lyapunov exponents of $f$ are 0 or none of the Lyapunov exponents of $f$ are 0 ($\mu$-a.e.)?

For some partially hyperbolic diffeomorphisms zero exponents were perturbed away in [ShWi2], which produces pathological center foliations. More of this is carried out in [HiPe] and for the $C^1$ topology in [BaBo]. So there is some evidence that the answer to the problem is "yes" at least for stably ergodic or partially hyperbolic diffeomorphisms. The problem is even interesting when restricted to ergodic diffeomorphisms so there is only one ergodic component. When $r = 1$, Mañé and Bochi prove for two dimensional manifolds that generically all the exponents are zero or the diffeomorphism is Anosov [Mañ1], [Boc].

**3.2. Possible Extensions.** How might the Anosov–Hopf argument be transported from the category of volume preserving diffeomorphisms to most of $\mathrm{Diff}^r(M)$? and especially to the existence of SRB measures? Here we enter a more speculative realm. First we recall the definition of SRB measures and some suggestions from [ShWi2].

Given $f \in \mathrm{Diff}^r(M)$ (not necessarily preserving $\mu$) a closed $f$ invariant set $A \subset M$ and an $f$ invariant ergodic measure $\nu$ on $A$, we define the *basin* of $A$ to be the set of points $x \in M$ such that $f^n(x) \to A$ and for every continuous function $\phi\colon M \to \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n}(\phi(x) + \cdots + \phi(f^n(x))) \to \int_A \phi(x)\, d\nu.$$

$\nu$ is an *SRB measure* and $A$ an *SRB attractor* (or an *ergodic attractor*) if the $\mu$ measure of the basin of $A$ with respect to $\mu$ is positive.

It follows from the definition that a diffeomorphism has at most countably many SRB measures. We can more formally describe the Sinai, Ruelle and Bowen [Si],

[Ru1], [BoRu] result already referred to. If $f$ is a $C^2$ Axiom A no cycle diffeomorphism then $\mu$ almost every point in $M$ is in a basin of an SRB measure and there are only finitely many SRB measures. It is this result that one would like to extend into a (more) general context.

The next problem obviously presents itself from the presentation.

**Problem 5.** For $r \geq 2$ is it true for the generic $f$ in $\mathrm{Diff}^r(M)$ that the union of the basins of the SRB attractors of $f$ has full $\mu$ measure in $M$?

This natural question is on the minds of quite a few people. See for example [Ru1], [BuPuShWi], [Pa], [Vi1], [You1], [You2], [BoDíVi]. One way to approach the problem along the lines of the Anosov–Hopf argument and as in [Pe1], [PuSh1] might be via an analogue of the either/or question on Lyapunov exponents for volume preserving diffeomorphisms above . For $r \geq 2$ there is no known analogue without the volume preserving hypothesis of the robust positive measure set of invariant tori with zero Lyapunov exponents which occurs via KAM theory. See [Vi2], [BuPuShWi].

**Problem 6.** For $r \geq 2$ is it true for the generic $f$ in $\mathrm{Diff}^r(M)$ and any weak limit $\nu$ of the push forwards $\frac{1}{n}\sum_1^n f_*^j\mu$ that almost every ergodic component of $\nu$ has some exponents not equal to 0 ($\nu$-a.e.)? All exponents not equal to 0?

Partially hyperbolic systems are a natural domain to begin considering problems [5] and [6]. When the volume is not preserved and we distinguish future behavior from the past the accessibility equivalence relation has to be adapted. Even for partially hyperbolic $f$ it is not entirely clear how to do this. So suppose $f$ partially hyperbolic . Let $W^{\mathrm{uu}}(x)$ and $W^{\mathrm{ss}}(x)$ denote the strong unstable and stable manifolds which are known to exist tangent to the $E^{\mathrm{u}}$ and $E^{\mathrm{s}}$ foliations. For $x, y \in M$ define $x > y$ if $W^{\mathrm{uu}}(x) \cap W^{\mathrm{ss}}(y) \neq \emptyset$. Transitivize $>$ to a partial order on $M$ and declare $x \sim y$ if $x > y$ and $y > x$. The $\sim$ equivalence classes may play a role similar to us-accessibility classes.

**Problem 7.** For the generic partially hyperbolic $f$, do all $\sim$ equivalence classes which are minimal with respect to $>$ have SRB measures?

**3.3. A little hyperbolicity.** Now that we have given a lot of examples, we return to our theme that a little hyperbolicity goes a long way towards ergodicity. We ask how often can we prove that hyperbolicity does exist in the guise of some non-zero Lyapunov exponents. Some specific families of non-uniformly hyperbolic dynamical systems have been worked out which contain positive measure sets in the parameter space with SRB measures having non-zero Lyapunov exponents. Most prominent among these families are the quadratic and Henon families, see [Ja], [Ly], [Be], [BeCa], [You1], [You2], [Vi1]. The proofs are difficult. One would like to find a fairly general principle which guarantees that a family has a positive measure set of parameters which have an SRB measure with a positive Lyapunov exponent.

One attempt posits that rich enough families of dynamical systems should have members with positive Lyapunov exponents. Examples have been constructed with uncertain but evocative results. Let $M$ have a Riemannian metric and let $G$ be a group of isometries of $M$ which is transitive on the projectivized tangent bundle of $M$. Let $\mu$ be the Riemannian volume. Let $f_\epsilon$ be a family of $C^r$ dynamical systems defined on $M$ depending on $\epsilon$. For fixed $\epsilon$, consider the family $Gf_\epsilon = \{gf_\epsilon, g \in G\}$. Give $Gf_\epsilon$ the push forward of the Haar measure on $G$. If $f_\epsilon$ preserves $\mu$ let $H(\epsilon)$ be the average over $Gf_\epsilon$ of the entropy of $gf_\epsilon$ with respect to $\mu$. This is the case in example 3 below. If $f_\epsilon$ does not preserve $\mu$ but $gf_\epsilon$ has a unique SRB measure for each $g \in G$, let $H(\epsilon)$ be the average over $Gf_\epsilon$ of the entropy of $gf_\epsilon$ with respect to this SRB measure. This is the case in examples 1 and 2 below. We compare $H(\epsilon)$ to the random Lyapunov exponents with respect to random products of elements of $Gf_\epsilon$ which we shall call

$$R(\epsilon) = \int_{\mathbb{P}TM} \ln|Tf_\epsilon(v)|\, dv,$$

where $\mathbb{P}TM$ is the projectivized tangent bundle of $M$. It is usually easy to see that $R(\epsilon)$ is positive. When $H(\epsilon)$ is positive then there are obviously positive measure sets in the parameter space with positive Lyapunov exponents and positive entropy. Here are the results for a few families.

1. *Blaschke products* ([PujRoSh]). The family of dynamical systems does not depend on $\epsilon$; we take $f_\epsilon = B$ where

$$B(z) = \theta_0 \prod_{i=1}^{n} \frac{z - a_i}{1 - z\overline{a_i}},$$

$n \geq 2$, $a_i \in \mathbb{C}$, $|a_i| < 1$, $i = 1, \ldots, n$, and $\theta_0 \in \mathbb{C}$ with $|\theta_0| = 1$.

The group $G$ is the unit circle $\mathbb{T}$ in the complex plane, $\mathbb{C}$. Its elements are denoted by $\theta$. Now we take

$$\mathbb{T}B = \{\theta B\}_{\{\theta \in \mathbb{T}\}}.$$

Then

$$H(\epsilon) \geq R(\epsilon).$$

($H(\epsilon)$ is always positive.)

2. *Expanding maps of the circle* ([LlShSi]). Here the dynamical systems are $f_{k,\alpha,\epsilon} : \mathbb{S}^1 \to \mathbb{S}^1$ which when written mod 1 are of the form

$$f_{k,\alpha,\epsilon} : x \mapsto kx + \alpha + \epsilon \sin(2\pi x). \tag{3.1}$$

The group is $\mathbb{S}^1$, $\alpha$ ranges over $\mathbb{S}^1$ and $k \geq 2$. Then for small $\epsilon$ the average over $\alpha$ of the entropy $H(\epsilon)$ is smaller than $R(\epsilon)$, while the max over $\alpha$ of the entropies of $f_{k,\alpha,\epsilon}$ is larger than $R(\epsilon)$. In the case of the averages the difference is on the order of $\epsilon^{2k+2}$. $H(\epsilon)$ is again obviously positive.

3. *Twist maps of the sphere* ([LeShSiWi]). For $\epsilon > 0$, we define a one-parameter family of twist maps $f_\epsilon$ as follows. Express $\mathbb{S}^2$ as the sphere of radius $1/2$ centered at $(0, 0)$ in $\mathbb{R} \times \mathbb{C}$, so that the coordinates $(r, z) \in \mathbb{S}^2$ satisfy the equation
$$|r|^2 + |z|^2 = 1/4.$$
In these coordinates define a twist map $f_\epsilon : \mathbb{S}^2 \to \mathbb{S}^2$, for $\epsilon > 0$, by
$$f_\epsilon(r, z) = (r, \exp(2\pi i (r + 1/2)\epsilon)z).$$
The group is SO(3). So $\mathrm{SO}(3) f_\epsilon = \{g f_\epsilon, g \in \mathrm{SO}(3)\}$.
For small $\epsilon$, $H(\epsilon)$ seems experimentally to be positive and is provably less than $R(\epsilon)$. $R(\epsilon)$ tends to infinity with $\epsilon$ and experimentally $R(\epsilon)$ and $H(\epsilon)$ are asymptotic. If we add a small fixed amount of randomization $\delta$ to the each $g$ in $g f_\epsilon$ and average the Lyapunov exponents of this randomized family over $g \in \mathrm{SO}(3)$, we obtain $R_\delta(\epsilon)$ which is indeed asymptotic to $R(\epsilon)$ as $\epsilon \to \infty$.

4. *Linear maps* ([DeSh]). If, instead of dynamical systems, we consider a linear map $A \in \mathrm{GL}(n, \mathbb{C})$ and the family $\mathrm{SU}(n)A$, then the average of the logarithms of the $k$ biggest moduli of eigenvalues of $UA$ over $U \in \mathrm{SU}(n)$ is greater than or equal to the sum of the $k$ largest Lyapunov exponents of random products of matrices from $\mathrm{SU}(n)A$.

There may be a general principle operating here that we have not put our finger on yet.

## 4. *Some* differentiable dynamical systems

**4.1. Affine diffeomorphisms.** The ergodic theory of affine diffeomorphisms of homogeneous spaces has been much studied in its own right, see for example [St1], and contains some of the principal examples studied in smooth dynamics such as the geodesic and horocycle flows on surfaces of constant negative curvature and toral automorphisms. Here we study the question of ergodicity of affine diffeomorphisms in the context of partially hyperbolic dynamical systems with $C^r$ perturbations. Our methods of proof recover the stable ergodicity of affine diffeomorphisms when they are stably ergodic among affine perturbations and usually extend this stability to $C^r$ perturbations. On this last point there remain some open problems.

Suppose that $G$ is a connected Lie group, $A : G \to G$ is an automorphism, $B$ is a closed subgroup of $G$ with $A(B) = B$, $g \in G$ is given, and the *affine diffeomorphism*
$$f : G/B \to G/B$$
is defined as $f(xB) = gA(x)B$. It is covered by the diffeomorphism
$$\bar{f} = L_g \circ A : G \to G,$$

where $L_g : G \to G$ is left multiplication by $g$.

An affine diffeomorphism $\bar{f}$ induces an automorphism of the Lie algebra $\mathfrak{g} = T_eG$, $\mathfrak{a}(\bar{f}) = \mathrm{Ad}_g \circ T_eA$, where $\mathrm{Ad}_g$ is the adjoint action of $g$, and $\mathfrak{g}$ splits into generalized eigenspaces,

$$\mathfrak{g} = \mathfrak{g}^{\mathrm{u}} \oplus \mathfrak{g}^{\mathrm{c}} \oplus \mathfrak{g}^{\mathrm{s}},$$

such that the eigenvalues of $\mathfrak{a}(\bar{f})$ are respectively outside, on, or inside the unit circle. These eigenspaces and the direct sums $\mathfrak{g}^{\mathrm{cu}} = \mathfrak{g}^{\mathrm{u}} \oplus \mathfrak{g}^{\mathrm{c}}$, $\mathfrak{g}^{\mathrm{cs}} = \mathfrak{g}^{\mathrm{c}} \oplus \mathfrak{g}^{\mathrm{s}}$ are Lie subalgebras and hence tangent to connected subgroups $G^{\mathrm{u}}$, $G^{\mathrm{c}}$, $G^{\mathrm{s}}$, $G^{\mathrm{cu}}$, $G^{\mathrm{cs}}$.

**Proposition** ([PuShSt1]). *Let $f : G/B \to G/B$ be an affine diffeomorphism as above such that $G/B$ is compact and supports a smooth $G$-invariant volume. Let $G^*$ be any of the groups $G^{\mathrm{u}}$, $G^{\mathrm{c}}$, $G^{\mathrm{s}}$, $G^{\mathrm{cu}}$, $G^{\mathrm{cs}}$. Then the orbits of the left $G^*$-action on $G/B$ foliate $G/B$. Moreover, $f$ exponentially expands the $G^{\mathrm{u}}$-leaves, exponentially contracts the $G^{\mathrm{s}}$-leaves, and affects the $G^{\mathrm{c}}$-leaves subexponentially.*

Now we characterize partial hyperbolicity, bunching and accessibility in the context of affine diffeomorphisms. Let $\mathfrak{h}$ denote the smallest Lie subalgebra of $\mathfrak{g}$ containing $\mathfrak{g}^{\mathrm{u}} \cup \mathfrak{g}^{\mathrm{s}}$. It is not hard to see that $\mathfrak{h}$ is an ideal in $\mathfrak{g}$. We call it the *hyperbolic Lie subalgebra* of $\bar{f}$, and we denote by $H$ the connected subgroup of $G$ tangent to $\mathfrak{h}$, calling it the *hyperbolic subgroup* of $\bar{f}$. Finally, let $\mathfrak{b}$ denote the Lie algebra of $B$, $\mathfrak{b} \subset \mathfrak{g}$.

**Theorem** ([PuSh4]). *Let $f : G/B \to G/B$ be an affine diffeomorphism as above such that $G/B$ is compact and supports a smooth $G$-invariant volume. Then*

(a) *$f$ is partially hyperbolic if and only if the hyperbolic Lie subalgebra of $\bar{f}$ is not contained in the Lie algebra of $B$, $\mathfrak{h} \not\subset \mathfrak{b}$.*

(b) *If $f$ is partially hyperbolic then it is center bunched.*

(c) *$f$ has the us-accessibility property if and only if $\mathfrak{g} = \mathfrak{b} + \mathfrak{h}$.*

(d) *$f$ has the e-accessibility property if and only if $\overline{HB} = G$.*

When the stable and unstable foliations are smooth, as in the affine case, us-accessibility is stable. Thus we have:

**Theorem** ([PuSh4]). *Let $f : G/B \to G/B$ be an affine diffeomorphism as above such that $G/B$ is compact and supports a smooth $G$-invariant volume. Then $f$ is stably ergodic among $C^2$ volume preserving diffeomorphisms of $G/B$ if* (*merely*) *the hyperbolic Lie subalgebra $\mathfrak{h}$ is large enough that $\mathfrak{g} = \mathfrak{b} + \mathfrak{h}$.*

If $G$ is simple then any nontrivial $\mathfrak{h}$ is large enough since it is an ideal.

Suppose that $A \in \mathrm{SL}(n, \mathbb{R})$ has some eigenvalues that are not of modulus one, and suppose that $\Gamma$ is a uniform discrete $A$-invariant subgroup of $\mathrm{SL}(n, \mathbb{R})$. Set $M = \mathrm{SL}(n, \mathbb{R})/\Gamma$. Then left multiplication by $A$, $L_A : M \to M$, is stably ergodic in

$\mathrm{Diff}^2_\mu(M)$. The case where $n$ is large and all but two eigenvalues have modulus one is interesting, in that the dimension of $G^{\mathrm{u}}$ and $G^{\mathrm{s}}$ is $n-1$ while the dimension of $G^{\mathrm{c}}$ is $(n-1)^2$, so the dimension of $G^{\mathrm{c}}$ is much larger than that of $G^{\mathrm{u}}$ and $G^{\mathrm{s}}$.

At the other extreme are abelian groups. If $G = \mathbb{R}^n$ and $B = \mathbb{Z}^n$ then translations on the torus, $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ are ergodic if the entries of the element defining the translation are rationally independent, but they are never stably ergodic. An automorphism $A$ of $\mathbb{T}^n$ is ergodic if and only if $A$ has no eigenvalues that are roots of unity. A little bit of algebra quickly shows that the hypothesis that $A$ has no eigenvalues which are roots of unity is equivalent to the hypothesis that $\overline{H\mathbb{Z}^n} = \mathbb{R}^n$ where $H$ is the hyperbolically generated subgroup of $\mathbb{R}^n$.

We have concentrated on the accessibility condition because accessibility is a topological property and as such it is not difficult to stipulate easily verifiable conditions which guarantee that it persists under small perturbations.

In a recent remarkable paper, Federico Rodriguez Hertz gives the first examples of a stably e-accessible diffeomorphisms that are not us-accessible, [RH]. They are ergodic, non-hyperbolic diffeomorphisms of tori. The first such occurs in dimension four.

Rodriguez Hertz sometimes uses a technical assumption on the automorphism $A$, which we will refer to as the *Rodriguez Hertz condition*, namely that the characteristic polynomial of $A$ is irreducible over the integers and it can not be written as a polynomial in $t^k$, $k \geq 2$.

**Theorem** ([RH]). *Let $A$ be an ergodic toral automorphism of $\mathbb{T}^n$.*

(a) *If $n \leq 5$ then $A$ is stably ergodic in* $\mathrm{Diff}^{22}_\mu(\mathbb{T}^n)$.

(b) *If $n \geq 6$, $E^{\mathrm{c}}$ is two-dimensional, and $A$ satisfies the Rodriguez Hertz condition then $A$ is stably ergodic in* $\mathrm{Diff}^5(\mathbb{T}^n)$.

The differentiability degrees 22 and 5 are not misprints.

Part of Rodriguez Hertz' proof involves an alternative. Either the perturbation is us-accessible or the stable and unstable manifold foliations are differentiably conjugate to the foliations of the linear example and hence the perturbation has the e-accessibility property.

**Problem 8.** Is every ergodic toral automorphism stably ergodic in the $C^r$ topology for some $r$?

The next result is an approximate solution of this problem.

**Theorem** ([ShWi1]). *Every ergodic toral automorphism of $\mathbb{T}^n$ that is an isometry on the center bundle $E^{\mathrm{c}}$ can be approximated arbitrarily well in* $\mathrm{Diff}^\infty_\mu(\mathbb{T}^n)$ *by a stably us-accessible, stably ergodic diffeomorphism.*

Further examples of partially hyperbolic stably ergodic diffeomorphisms are considered in [BuPuShWi]. These include skew products, frame flows, and Anosov-like diffeomorphisms.

The next theorem shows that the condition for stable ergodicity of affine diffeomorphisms among perturbations which are restricted to be left multiplication by group elements near the identity is the same as e-accessibility. Hence, the julienne proof of stable ergodicity applies to prove the stable ergodicity of these affine diffeomorphisms among affine perturbations as well. This phenomenon is discussed in [PuShSt2].

**Theorem** ([St2]). *Suppose that $f: G/B \to G/B$ is an affine diffeomorphism such that $M = G/B$ is compact and supports a smooth $G$-invariant volume. Then the following are equivalent.*

(a) *$f$ is stably ergodic under perturbation by left translations.*

(b) *$\overline{HB} = G$ where $H$ is the hyperbolically generated subgroup of $G$.*

**Corollary.** *Suppose that $f: G/B \to G/B$ is an affine diffeomorphism such that $M = G/B$ is compact and supports a smooth $G$-invariant volume. Assume that $G$ is simple. Then stable ergodicity of $f$ with respect to perturbation by left translations is equivalent to stable ergodicity among $C^2$ volume preserving perturbations.*

This corollary and the result of Rodriguez Hertz naturally lead to a generalization of Problem 10.

**Problem 9.** For an affine diffeomorphism $f$ of a compact, finite volume $G/B$, is stable ergodicity of $f$ with respect to perturbation by left translations equivalent to stable ergodicity among $C^2$ volume preserving perturbations?

We end our discussion of partially hyperbolic diffeomorphisms with a question from [BuPuShWi] of a very different nature. We have used both the strong unstable and strong stable foliations in our proof of ergodicity, but we do not know an example where this is strictly necessary.

**Problem 10.** For a partially hyperbolic $C^2$ ergodic diffeomorphism $f$ with the e-accessibility property, are the unstable and stable foliations already ergodic and uniquely ergodic?

Unique ergodicity of for horocycle flows was proved by Furstenberg [Fu]. Bowen and Marcus [BoMa] proved the unique ergodicity of the strong stable and unstable manifold foliations in the case where $f$ is the time-one map of a hyperbolic flow. Rodriguez Hertz' result adds more cases in which the invariant foliations are uniquely ergodic, namely those in which they are differentiably conjugate to the invariant foliations of a linear ergodic toral automorphism. Starkov [PuShSt2] proves that unique ergodicity of the strong stable or unstable foliations for all affine diffeomorphisms which are stably ergodic under perturbation by left translation.

In the topological category Bonatti, Díaz, and Ures [BoDíUr] prove the minimality of the stable and unstable foliations for an open and dense set of robustly transitive diffeomorphisms.

**4.2. Models.** Two dynamical systems $f\colon M \to M$ and $g\colon N \to N$ are topologically conjugate if there is a homeomorphism $h\colon M \to N$ such that $hf = gh$. The dynamical system $f$ is structurally stable if there is a $C^r$ neighborhood of $f$ such that every $g$ in $U$ is topologically conjugate to $f$.[¶] By the work of Smale, Palis [PaSm], Robbin [Ro] and Robinson [Rob], diffeomorphisms that satisfy Smale's Axiom A and the strong transversality condition are structurally stable. Mañé [Mañ2] in general and Liao [Li] also in dimension 2 prove that in the $C^1$ topology this condition is also necessary. The $C^2$ Axiom A strong transversality diffeomorphisms also have finitely many attractors which have SRB measures. These Axiom A strong transversality diffeomorphisms are extraordinarily appealing since they have all the properties we hope for. They are fairly well understood. Yet there remain interesting questions about them. Some of the issues are discussed in [Su]. I will denote the set of Axiom A strong transversality diffeomorphisms of $M$ by $\mathrm{AS}^r(M)$.[‖]

Since topological entropy is a topological conjugacy invariant and $C^\infty$ is dense in $C^r$ the Entropy Inequality holds for all $C^r$ structurally stable diffeomorphisms. How sharp is the Entropy Inequality as a lower bound for the entropy of dynamical systems in $\mathrm{AS}^r(M)$? Smale [Sm3] proved that every isotopy class of diffeomorphisms contains an element of $\mathrm{AS}^r(M)$. Since the fundamental group can contribute information about the entropy not readable in the homology groups, we restrict ourselves to simply connected manifolds.[**]

**Problem 11.** Let $M$ be simply connected. Let $I$ be an isotopy class of diffeomorphisms of $M$. Is there a sequence of diffeomorphisms, $f_n \in I \cap \mathrm{AS}^r(M)$ such that $h(f_n) \to \ln(s(f_*))$?

If the restriction that the diffeomorphism lie in $\mathrm{AS}^r(M)$ is removed then it is even unknown whether equality may be achieved in the Entropy Inequality within every isotopy class of diffeomorphisms. There are examples where equality may not be achieved with elements of $\mathrm{AS}^r(M)$. A diffeomorphism in $\mathrm{AS}^r(M)$ with zero entropy is necessarily Morse–Smale. As a result of [ShSu2], [FrSh] and [Le], it is known that there are isotopy classes of diffeomorphisms of simply connected manifolds for which $\ln(s(f_*)) = 0$, yet there is no Morse–Smale diffeomorphism in the class. Are there diffeomorphisms in $\mathrm{AS}^r(M)$ with arbitrarily small topological entropy in these classes? If not, what is a lower bound on the entropy?

Model elements of $\mathrm{AS}^r(M)$ are constructed in every isotopy class of diffeomorphisms in [ShSu2], [Fr2], [Mal1] from information on chain complexes for $M$ and chain complex endomorphisms induced by $f$. This work is closely related to Smale's work on the structure of manifolds. See also [Sh2], [Sh3] for more discussion of this point. There are further relations between stability and homology theory established

[¶] We restrict ourselves to dynamical systems in $\mathrm{Diff}^r(M)$ even though the same concepts apply in $\mathrm{End}^r(M)$ and to structural stability as opposed to Omega stability for the sake of simplicity of exposition.

[‖] AS is a fortuitous selection of letters since Anosov, Sinai, Smale, Axiom A and Strong all begin with A and S.

[**] See Maller [Mal1], [Mal2] for non-simply connected manifolds.

in [ShWil], [RuSu] where the entropy conjecture was first proven for $C^1$ diffeomorphisms satisfying Smale's axioms. This work is also related to our next problem.

To close our discussion of structurally stable diffeomorphisms, I recall one other outstanding problem.

**Problem 12.** Are all Anosov diffeomorphisms infra-nil?

Smale [Sm2], considered the nil-manifold setting for Anosov diffeomorphisms which was later extended by example [Sh1] to infra-nil manifolds where the corresponding examples of expanding maps were considered. All expanding maps are infra-nil by the results of [Sh1], [Fr1] and Gromov [Gro1] on groups of polynomial growth. For Anosov diffeomorphisms defined on a manifold $M$, it is known that if $M$ is an infra-nil manifold then the diffeomorphism is conjugate to an affine example, [Ma2]. It is not known if all manifolds $M$ supporting Anosov diffeomorphisms are infra-nil manifolds. If one of the bundles $E^{\mathrm{s}}$ or $E^{\mathrm{u}}$ is one dimensional then problem is answered in the affirmative by [New2]. Perhaps the best results go back to [Fr1].

Questions about the classification of manifolds admitting partially hyperbolic diffeomorphisms are raised in section 20 of [PuSh5].

We end the paper by mentioning a few surveys which go into greater depth on some of the issues we have considered, [Sm2], [Fr2], [Sh1], [Sh3], [BuPuShWi], [PuSh5], [Pe3], [BoDíVi].

## References


[An] Anosov, D. V., Geodesic flows on closed Riemannian manifolds of negative curvature. *Proc. Steklov. Inst. Math.* **90** (1967).

[AnSi] Anosov, D. V., and Sinai, Ya. G., Some smooth ergodic systems. *Russian Math. Surveys* **22** (5) (1967), 103–167.

[BaBo] Baraviera, A., and Bonatti, C., Removing zero Lyapunov exponents. *Ergodic Theory Dynam. Systems* **23** (2003), 1655–1670.

[Be] Benedicks, M., Non uniformly hyperbolic dynamics: Hénon maps and related dynamical systems. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, 255–264; Errata, Vol. I, Higher Ed. Press, Beijing 2002, 651.

[BeCa] Benedicks, M., and Carleson, L., The dynamics of the Henon map. *Ann. of Math.* **133** (1991), 73–169.

[Boc] Bochi, J., Genericity of zero Lyapunov exponents. *Ergodic Theory Dynam. Systems* **22** (2002), 1667–1696.

[BoDíUr] Bonatti, C., Díaz, L. J., Ures, R., Minimality of strong stable and unstable foliations for partially hyperbolic diffeomorphisms. *J. Inst. Math. Jussieu* **1** (2002), 513–541.

[BoDíVi] Bonatti, C., Díaz, L.J., Viana, M., *Dynamics Beyond Uniform Hyperbolicity*. Encyclopedia of Mathematical Sciences 102, Mathematical Physics III, Springer-Verlag, Berlin 2005.

[Bo] Bowen, R., Entropy and the fundamental group. In *The structure of attractors in dynamical systems* (Proc. Conf., North Dakota State Univ., Fargo, N.D., 1977), Lecture Notes in Math. 668, Springer-Verlag, Berlin 1978, 21–29.

[BoMa] Bowen, R., and Marcus, B., Unique ergodicity for horocycle foliations. *Israel J. Math.* **26** (1977), 43–67.

[BoRu] Bowen, R., and Ruelle, D., The ergodic theory of Axiom A flows. *Invent. Math.* **29** (1975), 181–202.

[Br1] Brin, M., Topological transitivity of one class of dynamical systems and flows of frames on manifolds of negative curvature. *Funct. Anal. Appl.* **9** (1975), 8–16.

[Br2] Brin, M., The topology of group extensions of $C$ systems. *Mat. Zametki* **18** (1975), 453–465.

[BrPe] Brin, M., and Pesin, Ja., Partially hyperbolic dynamical systems. *Math. USSR Izvestija* **8** (1974), 177–218.

[BuPuShWi] Burns, K., Pugh, C., Shub, M., and Wilkinson, A., Recent results about stable ergodicity. In *Smooth ergodic theory and its applications*, Proc. Sympos. Pure Math. 69, Amer. Math. Soc., Providence, RI, 2001, 327–366.

[BuWi1] Burns, K., and Wilkinson, A., Stable ergodicity of skew products. *Ann. Sci. École Norm. Sup.* **32** (1999), 859–889.

[BuWi2] Burns, K., and Wilkinson, A., Better center bunching. Preprint.

[BuWi3] Burns, K., and Wilkinson, A., On the ergodicity of partially hyperbolic systems. Preprint.

[ChSu] Cheng, C.-Q., and Sun, Y.-S., Existence of invariant tori in three dimensional measure-preserving mappings. *Celestial Mech. Dynam. Astronom.* **47** (1989/90), 275–292.

[DeSh] Dedieu, J.-P., and Shub, M., On Random and Mean Exponents for Unitarily Invariant Probability Measures on $GL_n(\mathbb{C})$. In *Geometric methods in dynamics* (II), *Asterisque* **287** (2003), 1–18.

[LlShSi] de la Llave, R., Shub, M., and Simó, C., Entropy estimates for a family of expanding maps of the circle. Preprint.

[DoWi] Dolgopyat, D., and Wilkinson, A., Stable accessibility is $C^1$ dense. *Astérisque* **287** (2003), 33–60.

[FaSh] Fathi, A., and Shub, M., Some Dynamics of Pseudo-Anosov Diffeomorphisms. In *Travaux de Thurston sur les Surfaces*, *Asterisque* **66–67** (1979), 181–207.

[FiPa] Field, M., and Parry, W., Stable ergodicity of skew extensions by compact Lie groups. *Topology* **38** (1999), 167–187.

[Fr1] Franks, J. M., Anosov Diffeomorphisms. In *Global Analysis*, Proc. Sympos. Pure Math. 14, Amer. Math. Soc., Providence, R.I., 1970, 61–93.

[Fr2] Franks, J. M., *Homology Theory and Dynamical Systems*. CBMS Reg. Conf. Ser. Math. 49, Amer. Mat. Soc., Providence, RI, 1982.

[FrSh] Franks, J., and Shub, M., The Existence of Morse-Smale Diffeomorphisms. *Topology* **20** (1981), 273–290.

[Fu] Furstenberg, H., The unique ergodicity of the horocycle flow. In *Recent advances in topological dynamics* (Proc. Conf., Yale Univ., New Haven, Conn., 1972; in honor of Gustav Arnold Hedlund), Lecture Notes in Math. 318, Springer-Verlag, Berlin 1973, 95–115.

[GrPuSh] Grayson, M., Pugh, C., and Shub, M., Stably ergodic diffeomorphisms. *Ann. of Math.* (2) **140** (1994), 295–329.

[Gro1] Gromov, M., Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[Gro2] Gromov, M., Entropy, homology and semialgebraic geometry. In *Séminaire Bourbaki*, Vol. 1985/86; *Astérisque* **145–146** (5) (1987), 225–240

[HiPe] Hirayama, M., and Pesin, Ya., Non-absolutely Continuous Foliations. Preprint.

[Ho] Hopf, E., Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. *Ber. Verh. Sächs. Akad. Wiss. Leipzig* **91** (1939), 261–304.

[Ja] Jakobson, M., Absolutely continuous invariant measures for one-parameter families of one dimensioanl maps. *Comm. Math. Phys.* **81** (1981), 39–88.

[Ka] Katok, A., Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst. Hautes Études Sci. Publ. Math.* **51** (1980), 137–173.

[LeShSiWi] Ledrappier, F., Shub, M., Simó, C., and Wilkinson, A., Random versus deterministic exponents in a rich family of diffeomorphisms. *J. Stat. Phys.* **113** (2003), 85–149.

[Le] Lenstra, H. W. Jr., Grothendieck groups of abelian group rings. *J. Pure Appl. Algebra* **20** (1981), 173–193.

[Li] Liao, S. T., On the stability conjecture. *Chinese Ann. Math.* **1** (1980), 9–30.

[Ly] Lyubich, M., Almost every real quadratic map is either regular or stochastic. *Ann. of Math.* (2) **156** (2002), 1–78.

[Mal1] Maller, M., Fitted diffeomorphisms of nonsimply connected manifolds. *Topology* **19** (1980), 395–410.

[Mal2] Maller, M., Algebraic problems arising from Morse-Smale dynamical systems. In *Geometric dynamics* (Rio de Janeiro, 1981), Lecture Notes in Math. 1007, Springer-Verlag, Berlin 1983, 512–521.

[Ma1] Manning, A., Topological entropy and the first homology group. In *Dynamical systems—Warwick 1974* (Proc. Sympos. Appl. Topology and Dynamical Systems, Univ. Warwick, Coventry, 1973/1974; presented to E. C. Zeeman on his fiftieth birthday), Lecture Notes in Math. 468, Springer-Verlag, Berlin 1975, 185–190.

[Ma2] Manning, A., There are no new Anosov diffeomorphisms on tori. *Amer. J. Math.* **96** (1974), 422–429.

[Mañ1] Mañé, R., Oseledec's theorem from the generic viewpoint. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 2, PWN, Warsaw 1984, 1269–1276.

[Mañ2] Mañé, R., A proof of the $C^1$ stability conjecture. *Inst. Hautes Études Sci. Publ. Math.* **66** (1988), 161–210.

[MaPr] Marzantowicz, W., and Przytycki, F., Entropy conjecture for continuous maps of nilmanifolds. Preprint.

[MiPr] Misiurewicz, M., and Przytycki, F., Topological entropy and degree of smooth mappings. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **25** (1977), 573–574.

[New1] Newhouse, S. E., Entropy in Smooth Dynamical Systems. In *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. II, The Mathematical Society of Japan, Tokyo, Springer-Verlag, Tokyo, 1991, 1285–1294.

[New2] Newhouse, S. E., On codimension one Anosov diffeomorphisms. *Amer. J. Math.* **92** (1970), 761–770.

[Pa] Palis, J., A Global View of Dynamics and a Conjecture on the Denseness of Finitude of Attractors. *Astérisque* **261** (2000), 339–351.

[PaSm] Palis, J., and Smale, S., Structural stability theorems. In *Global Analysis* , Proc. Sympos. Pure Math. 14, Amer. Math. Soc., Providence, R.I., 1970, 223–231.

[Pe1] Pesin, Ya., Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Math. Surveys* **32** (4) (1977), 55–114.

[Pe2] Pesin, Ya., Ergodic properties and dimensionlike characteristics of strange attractors that are close to hyperbolic. In *Proceedings of the International Congress of Mathematicians* (Berkeley, Calif., 1986), Vol. 2, Amer. Math. Soc., Providence, R.I., 1987, 1195–1209.

[Pe3] Pesin, Ya., *Lectures on partial hyperbolicity and stable ergodicity*. Zurich Lectures in Advanced Mathematics, EMS Publishing House, Zürich 2004.

[PeSi] Pesin, Ya. B., and Sinai, Ya. G., Gibbs measures for partially hyperbolic attractors. *Ergodic Theory Dynam. Systems* **2** (1982), 417–438.

[Pu] Pugh, C., On the entropy conjecture: a report on conversations among R. Bowen, M. Hirsch, A. Manning, C. Pugh, B. Sanderson, M. Shub, and R. Williams. In *Dynamical systems—Warwick 1974* (Proc. Sympos. Appl. Topology and Dynamical Systems, Univ. Warwick, Coventry, 1973/1974; presented to E. C. Zeeman on his fiftieth birthday), Lecture Notes in Math. 468, Springer-Verlag, Berlin 1975, 257–261.

[PuSh1] Pugh, C., and Shub, M., Ergodic Attractors. *Trans. Amer. Math. Soc.* **312** (1989), 1–54.

[PuSh2] Pugh, C., and Shub, M., Stable Ergodicity and Partial Hyperbolicity. In *International Conference on Dynamical Systems: Montevideo 1995, a tribute to Ricado Mane* (ed. by F. Ledrappier, J. Lewowicz and S. Newhouse), Pitman Res. Notes Math. 362, Longman, Harlow 1996, 182–187.

[PuSh3] Pugh, C., and Shub, M., Stably ergodic dynamical systems and partial hyperbolicity. *J. Complexity* **13** (1997), 125–179.

[PuSh4] Pugh, C., and Shub, M., Stable ergodicity and julienne quasiconformality. *J. Eur. Math. Soc.* **2** (2000), 1–52.

[PuSh5] Pugh, C., and Shub, M., Stable ergodicity. *Bull. Amer. Math. Soc.* (N.S.) **41** (2004), 1–41.

[PuShSt1] Pugh, C., Shub, M., and Starkov, A., Stable ergodicity and julienne quasiconformality. *J. Eur. Math. Soc.* **2** (2000), 1–52; Corrigendum, ibid. **6** (2004), 149–151.

[PuShSt2] Pugh, C., Shub, M., and Starkov, A., Unique Ergodicity, Stable Ergodicity and the Mautner Phenomenon for Diffeomorhisms. *Discrete Contin. Dyn. Syst. Ser. A* **14** (2006), 845–855.

[Puj] Pujals, E., Tangent bundles dynamics and its consequences. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing 2002, 327–338.

[PujRoSh] Pujals, E., Robert, L., and Shub, M., Expanding maps of the circle rerevisited: Positive Lyapunov exponents in a rich family. Preprint.

[Ro] Robbin, J., A structural stability theorem. *Ann. of Math.* (2) **94** (1971), 447–493.

[Rob] Robinson, C., Structural Stability of $C^1$ diffeomorphisms. *J. Differential Equations* **22** (1976), 28–73.

[RH] Rodriguez Hertz, F., Stable Ergodicity of Certain Linear Automorphisms of the Torus. To appear.

[RHRHUr] Rodriguez Hertz, F., Rodriguez Hertz, J., and Ures, R., Partially hyperbolic systems with 1D-center bundle: I. Stable Ergodicity. Preprint

[Ru1] Ruelle, D., A measure associated with Axiom-A attractors. *Amer. J. Math.* **98** (1976), 619–654.

[Ru2] Ruelle, D., Turbulent Dynamical Systems. In *Proceedings International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 271–286.

[RuSu] Ruelle, D., and Sullivan, D., Currents, flows and diffeomorphisms. *Topology* **14** (1975), 319–327.

[RuWi] Ruelle, D., and Wilkinson, A., Absolutely singular dynamical foliations. *Comm. Math. Phys.* **219** (2001), 481–487.

[Sh1] Shub, M., Endomorphisms of Compact Differentiable Manifolds. *Amer. J. Math.* **91** (1969), 175–199.

[Sh2] Shub, M., Dynamical Systems, Filtrations and Entropy. *Bull. Amer. Math. Soc.* **80** (1974), 27–41.

[Sh3] Shub, M., The Geometry and Topology of Dynamical Systems and Algorithms for Numerical Problems. In *Proceedings of the 1983 Beijing Symposium on Differential Geometry and Differential Equations* (ed. by Liao Shantao), Science Press, Beijing 1986, 231–260.

[ShSu1] Shub, M., and Sullivan, D., A Remark on the Lefschetz Fixed Point Formula for Differentiable Maps. *Topology* **13** (1974), 189–191.

[ShSu2] Shub, M., and Sullivan, D., Homology Theory and Dynamical Systems. *Topology* **14** (1975), 109–132.

[ShWi1] Shub, M., and Wilkinson, A., Stably ergodic approximation: two examples. *Ergodic Theory Dynam. Systems* **20** (2000), 875–894.

[ShWi2] Shub, M., and Wilkinson, A., Pathological foliations and removable zero exponents. *Invent. Math.* **139** (2000), 495–508.

[ShWil] Shub, M., and Williams, R., Entropy and Stability. *Topology* **14** (1975), 329–338.

[Si] Sinai, Ya., Gibbs measures in ergodic theory. *Russian Math. Surveys* **27** (1972), 21–69.

[Sm1] Smale, S., On the structure of manifolds. *Amer. J. Math.* **84** (1962), 387–399.

[Sm2] Smale, S., Differentiable dynamical systems. *Bull. Amer. Math. Soc.* **73** (1967), 747–817.

[Sm3] Smale, S., Stability and isotopy in discrete dynamical systems. In *Dynamical Systems* (ed. by M. M. Peixoto), Academic Press, New York 1973, 527–530.

[St1] Starkov, A. N., *Dynamical Systems on Homogeneous Spaces*. Transl. Math. Monographs 190, Amer. Math. Soc., Providence, RI, 2000.

[St2] Starkov, A. N., Stable ergodicity among left translations. Appendix to Pugh, C., and Shub, M., Stable ergodicity and julienne quasiconformality. *J. Eur. Math. Soc.* **2** (2000), 1–52.

[Su] Sullivan, D., Inside and Outside Manifolds. In *Proceedings of the International Congress of Mathematicians, 1974* (Vancouver, B. C., 1974), Vol 1, Canad. Math. Congress, Montreal, Que., 201–208.

[Vi1] Viana, M., Dynamics: a probabilistic and geometric perspective. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. I, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 557–578.

[Vi2] Viana, M., Almost all cocycles over any hyperbolic system have non-vanishing Lyapunov exponents. Preprint

[Wi1] Wilkinson, A., Stable ergodicity of the time one map of a geodesic flow. Ph.D. Thesis, University of California at Berkeley, 1995.

[Wi2] Wilkinson, A., Stable ergodicity of the time one map of a geodesic flow. *Ergodic Theory Dynam. Systems* **18** (1998), 1545–1588.

[Xi] Xia, Z., Existence of invariant tori in volume-preserving diffeomorphisms. *Ergodic Theory Dynam. Systems* **12** (1992), 621–631.

[Yoc] Yoccoz, J.-C., Travaux de Herman sur les tores invariants. In *Séminaire Bourbaki*, Vol. 1991/92; *Astérisque* **206** (1992), **4**, 311–344.

[Yom] Yomdin, Y., Volume growth and entropy. *Israel J. Math.* **57** (1987), 285–300.

[You1] Young, L.-S., Ergodic Theory of Attractors. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 1230–1237.

[You2] Young, L.-S., What are SRB measures, and which dynamical systems have them? *J. Stat. Phys.* **108** (2002), 733–754.

Mathematics Department, University of Toronto, Bahen Centre, 40 St George St,
Room 6290, Toronto, Ontario M5S 2E4, Canada

E-mail: michael.shub@utoronto.ca

# Geodesics on flat surfaces

Anton Zorich*

**Abstract.** Various problems of geometry, topology and dynamical systems on surfaces as well as some questions concerning one-dimensional dynamical systems lead to the study of closed surfaces endowed with a flat metric with several cone-type singularities. In an important particular case, when the flat metric has trivial holonomy, the corresponding flat surfaces are naturally organized into families which appear to be isomorphic to moduli spaces of holomorphic one-forms.

One can obtain much information about the geometry and dynamics of an individual flat surface by studying both its orbit under the Teichmüller geodesic flow and under the linear group action on the corresponding moduli space. We apply this general principle to the study of generic geodesics and to counting of closed geodesics on a flat surface.

**Mathematics Subject Classification (2000).** Primary 57M50, 32G15; Secondary 37D40, 37D50, 30F30.

**Keywords.** Flat surface, Teichmüller geodesic flow, moduli space, asymptotic cycle, Lyapunov exponent, interval exchange transformation, renormalization.

## Introduction: families of flat surfaces as moduli spaces of Abelian differentials

Consider a collection of vectors $\vec{v}_1, \dots, \vec{v}_n$ in $\mathbb{R}^2$ and construct from these vectors a broken line in a natural way: a $j$-th edge of the broken line is represented by the vector $\vec{v}_j$. Construct another broken line starting at the same point as the initial one by taking the same vectors in the order $\vec{v}_{\pi(1)}, \dots, \vec{v}_{\pi(n)}$, where $\pi$ is some permutation of $n$ elements. By construction the two broken lines share the same endpoints; suppose that they bound a polygon like in Figure 1. Identifying the pairs of sides corresponding to the same vectors $\vec{v}_j$, $j = 1, \dots, n$, by parallel translations we obtain a surface endowed with a flat metric. (This construction follows the one in [M1].) The flat metric is nonsingular outside of a finite number of cone-type singularities corresponding to the vertices of the polygon. By construction the flat metric has trivial holonomy: a parallel transport of a vector along a closed path does not change the direction (and length) of the vector. This implies, in particular, that all cone angles are integer multiples of $2\pi$.

The polygon in our construction depends continuously on the vectors $\vec{v}_j$. This means that the combinatorial geometry of the resulting flat surface (its genus $g$, the

*The author is grateful to the MPI (Bonn) and to IHES for hospitality during the preparation of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

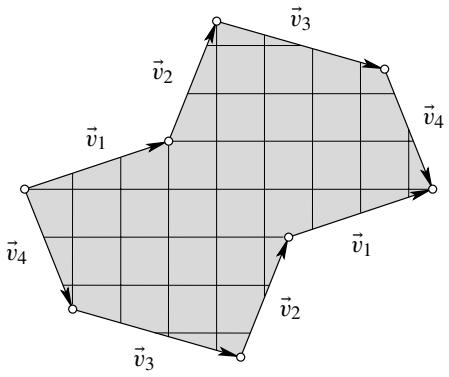



Figure 1. Identifying corresponding pairs of sides of this polygon by parallel translations we obtain a surface of genus two. It has single conical singularity with cone angle $6\pi$; the flat metric has trivial holonomy.

number $m$ and types of the resulting conical singularities) does not change under small deformations of the vectors $\vec{v}_j$. This allows to consider a flat surface as an element of a family of flat surfaces sharing common combinatorial geometry; here we do not distinguish isometric flat surfaces. As an example of such family one can consider a family of flat tori of area one, which can be identified with the space of lattices of area one:

$$\mathrm{SO}(2,\mathbb{R}) \backslash \mathrm{SL}(2,\mathbb{R}) / \mathrm{SL}(2,\mathbb{Z}) \quad = \quad \mathbb{H}^2 / \mathrm{SL}(2,\mathbb{Z})$$

The corresponding "modular surface" is not compact, see Figure 2. Flat tori representing points, which are close to the cusp, are almost degenerate: they have a very short closed geodesic. Similarly, families of flat surfaces of higher genera also form noncompact finite-dimensional orbifolds. The origin of their noncompactness is the same as for the tori: flat surfaces having short closed geodesics represent points which are close to the multidimensional "cusps".

We shall consider only those flat surfaces, which have trivial holonomy. Choosing a direction at some point of such flat surface we can transport it to any other point. It would be convenient to include the choice of direction in the definition of a flat structure. In particular, we want to distinguish the flat structure represented by the polygon in Figure 1 and the one represented by the same polygon rotated by some angle different from $2\pi$.

Consider the natural coordinate $z$ in the complex plane. In this coordinate the parallel translations which we use to identify the sides of the polygon in Figure 1 are represented as $z' = z + \mathrm{const}$. Since this correspondence is holomorphic, it means that our flat surface $S$ with punctured conical points inherits the complex structure. It is easy to check that the complex structure extends to the punctured points. Consider now a holomorphic 1-form $dz$ in the complex plane. When we pass to the surface $S$ the coordinate $z$ is not globally defined anymore. However, since the changes of

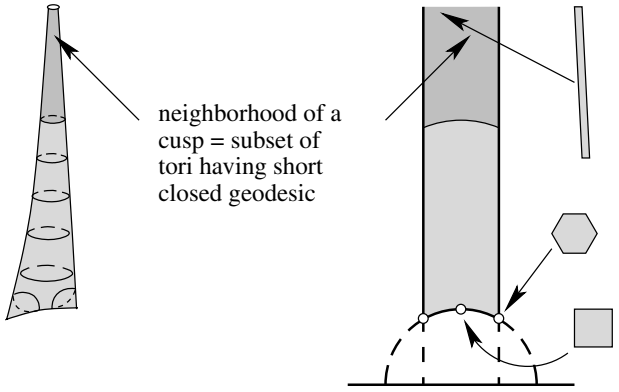



Figure 2. "Modular surface" $\mathbb{H}^2/\operatorname{SL}(2,\mathbb{Z})$ representing the space of flat tori is a noncompact orbifold of finite volume.

local coordinates are defined as $z' = z + \text{const}$, we see that $dz = dz'$. Thus, the holomorphic 1-form $dz$ on $\mathbb{C}$ defines a holomorphic 1-form $\omega$ on $S$ which in local coordinates has the form $\omega = dz$. It is easy to check that the form $\omega$ has zeroes exactly at those points of $S$ where the flat structure has conical singularities.

Reciprocally, one can show that a pair (Riemann surface, holomorphic 1-form) uniquely defines a flat structure of the type described above.

In an appropriate local coordinate $w$ a holomorphic 1-form can be represented in a neighborhood of zero as $w^d\,dw$, where $d$ is called the degree of zero. The form $\omega$ has a zero of degree $d$ at a conical point with cone angle $2\pi(d+1)$. The sum of degrees $d_1 + \cdots + d_m$ of zeroes of a holomorphic 1-form on a Riemann surface of genus $g$ equals $2g-2$. The moduli space $\mathcal{H}_g$ of pairs (complex structure, holomorphic 1-form) is a $\mathbb{C}^g$-vector bundle over the moduli space $\mathcal{M}_g$ of complex structures. The space $\mathcal{H}_g$ is naturally stratified by the strata $\mathcal{H}(d_1,\dots,d_m)$ enumerated by unordered partitions of the number $2g-2$ in a collection of positive integers $2g-2 = d_1+\cdots+d_m$. Any holomorphic 1-forms corresponding to a fixed stratum $\mathcal{H}(d_1,\dots,d_m)$ has exactly $m$ zeroes, and $d_1,\dots,d_m$ are the degrees of zeroes. Note, that an individual stratum $\mathcal{H}(d_1,\dots,d_m)$ in general does not form a fiber bundle over $\mathcal{M}_g$.

It is possible to show that if the permutation $\pi$ which was used to construct a polygon in Figure 1 satisfy some explicit conditions, vectors $\vec{v}_1,\dots,\vec{v}_n$ representing the sides of the polygon serve as coordinates in the corresponding family $\mathcal{H}(d_1,\dots,d_m)$. Consider vectors $\vec{v}_j$ as complex numbers. Let $\vec{v}_j$ join vertices $P_j$ and $P_{j+1}$ of the polygon. Denote by $\rho_j$ the resulting path on $S$ joining the points $P_j$, $P_{j+1} \in S$. Our interpretation of $\vec{v}_j$ as of a complex number implies that

$$\int_{\rho_j} \omega = \int_{P_j}^{P_{j+1}} dz = v_j \in \mathbb{C}.$$

The path $\rho_j$ represents a relative cycle: an element of the relative homology group

$H_1(S, \{P_1, \dots, P_m\}; \mathbb{Z})$ of the surface $S$ relative to the finite collection of conical points $\{P_1, \dots, P_m\}$. Relation above means that $\vec{v}_j$ represents a period of $\omega$: an integral of $\omega$ over the relative cycle $\rho_j$. In other words, a small domain in $H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$ containing $[\omega]$ can be considered as a local coordinate chart in our family $\mathcal{H}(d_1, \dots, d_m)$ of flat surfaces.

We summarize the correspondence between geometric language of flat surfaces and the complex-analytic language of holomorphic 1-forms on a Riemann surface in the dictionary below.

| Geometric language | Complex-analytic language |
|---|---|
| flat structure (including a choice of the vertical direction) | complex structure and a choice of a holomorphic 1-form $\omega$ |
| conical point with a cone angle $2\pi(d+1)$ | zero of degree $d$ of the holomorphic 1-form $\omega$ (in local coordinates $\omega = w^d\, dw$) |
| side $\vec{v}_j$ of a polygon | relative period $\int_{P_j}^{P_{j+1}} \omega = \int_{\vec{v}_j} \omega$ of the 1-form $\omega$ |
| family of flat surfaces sharing the same cone angles $2\pi(d_1+1), \dots, 2\pi(d_m+1)$ | stratum $\mathcal{H}(d_1, \dots, d_m)$ in the moduli space of Abelian differentials |
| coordinates in the family: vectors $\vec{v}_i$ defining the polygon | coordinates in $\mathcal{H}(d_1, \dots, d_m)$ : relative periods of $\omega$ in $H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$ |

Note that the vector space $H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$ contains a natural integer lattice $H^1(S, \{P_1, \dots, P_m\}; \mathbb{Z} \oplus \sqrt{-1}\,\mathbb{Z})$. Consider a linear volume element $d\nu$ in the vector space $H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$ normalized in such a way that the volume of the fundamental domain in the "cubic" lattice

$$H^1(S, \{P_1, \dots, P_m\}; \mathbb{Z} \oplus \sqrt{-1}\,\mathbb{Z}) \subset H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$$

equals one. Consider now the real hypersurface $\mathcal{H}_1(d_1, \dots, d_m) \subset \mathcal{H}(d_1, \dots, d_m)$ defined by the equation $area(S) = 1$. The volume element $d\nu$ can be naturally restricted to the hypersurface defining the volume element $d\nu_1$ on $\mathcal{H}_1(d_1, \dots, d_m)$.

**Theorem** (H. Masur; W. A. Veech). *The total volume* $\mathrm{Vol}(\mathcal{H}_1(d_1, \dots, d_m))$ *of every stratum is finite.*

The values of these volumes were computed by A. Eskin and A. Okounkov [EO].

Consider a flat surface $S$ and consider a polygonal pattern obtained by unwrapping $S$ along some geodesic cuts. For example, one can assume that our flat surface $S$ is glued from a polygon $\Pi \subset \mathbb{R}^2$ as on Figure 1. Consider a linear transformation $g \in \mathrm{GL}^+(2, \mathbb{R})$ of the plane $\mathbb{R}^2$. The sides of the new polygon $g\Pi$ are again arranged into pairs, where the sides in each pair are parallel and have equal length. Identifying the sides in each pair by a parallel translation we obtain a new flat surface $gS$ which, actually, does not depend on the way in which $S$ was unwrapped to a polygonal pattern $\Pi$. Thus, we get a continuous action of the group $\mathrm{GL}^+(2, \mathbb{R})$ on each stratum $\mathcal{H}(d_1, \dots, d_m)$.

Considering the subgroup $\mathrm{SL}(2, \mathbb{R})$ of area preserving linear transformations we get the action of $\mathrm{SL}(2, \mathbb{R})$ on the "unit hyperboloid" $\mathcal{H}_1(d_1, \dots, d_m)$. Considering the diagonal subgroup $\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \subset \mathrm{SL}(2, \mathbb{R})$ we get a continuous action of this one-parameter subgroup on each stratum $\mathcal{H}(d_1, \dots, d_m)$. This action induces a natural flow on the stratum which is called the *Teichmüller geodesic flow*.

**Key Theorem** (H. Masur; W. A. Veech). *The actions of the groups* $\mathrm{SL}(2, \mathbb{R})$ *and* $\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$ *preserve the measure* $d\nu_1$. *Both actions are ergodic with respect to this measure on each connected component of every stratum* $\mathcal{H}_1(d_1, \dots, d_m)$.

The following basic principle (which was first used in the pioneering works of H. Masur [M1] and of W. Veech [V1] to prove unique ergodicity of almost all interval exchange transformations) appeared to be surprisingly powerful in the study of flat surfaces. Suppose that we need some information about geometry or dynamics of an individual flat surface $S$. Consider the "point" $S$ in the corresponding family of flat surfaces $\mathcal{H}(d_1, \dots, d_m)$. Denote by $\mathcal{N}(S) = \overline{\mathrm{GL}^+(2, \mathbb{R})\, S} \subset \mathcal{H}(d_1, \dots, d_m)$ the closure of the $\mathrm{GL}^+(2, \mathbb{R})$-orbit of $S$ in $\mathcal{H}(d_1, \dots, d_m)$.

In numerous cases knowledge about the structure of $\mathcal{N}(S)$ gives a comprehensive information about geometry and dynamics of the initial flat surface $S$. Moreover, some delicate numerical characteristics of $S$ can be expressed as averages of simpler characteristics over $\mathcal{N}(S)$. We apply this general philosophy to the study of geodesics on flat surfaces.

Actually, there is a hope that this philosophy extends much further. A closure of an orbit of an abstract dynamical system might have extremely complicated structure. According to the optimistic hopes, the closure $\mathcal{N}(S)$ of a $\mathrm{GL}^+(2, \mathbb{R})$-orbit of any flat surface $S$ is a nice complex-analytic variety, and all such varieties might be classified. For genus two the latter statements were recently proved by C. McMullen (see [Mc1] and [Mc2]) and partly by K. Calta [Ca].

The following theorem supports the hope for some nice and simple description of orbit closures.

**Theorem** (M. Kontsevich). *Suppose that the closure in the stratum $\mathcal{H}(d_1, \dots, d_m)$ of a $\mathrm{GL}^+(2, \mathbb{R})$-orbit of some flat surface $S$ is a complex-analytic subvariety. Then in cohomological coordinates $H^1(S, \{P_1, \dots, P_m\}; \mathbb{C})$ this subvariety is represented by an affine subspace.*

## 1. Geodesics winding up flat surfaces

In this section we study geodesics on a flat surface $S$ going in generic directions. According to the theorem of S. Kerckhoff, H. Masur and J. Smillie [KeMS], for any flat surface $S$ the directional flow in almost any direction is uniquely ergodic. This implies, in particular, that for such directions the geodesics wind around $S$ in a relatively regular manner. Namely, it is possible to find a cycle $c \in H_1(S; \mathbb{R})$ such that a long piece of geodesic pretends to wind around $S$ repeatedly following this asymptotic cycle $c$. Rigorously it can be described as follows. Having a geodesic segment $X \subset S$ and some point $x \in X$ we emit from $x$ a geodesic transversal to $X$. From time to time the geodesic would intersect $X$. Denote the corresponding points as $x_1, x_2, \dots$. Closing up the corresponding pieces of the geodesic by joining the starting point $x_0$ and the point $x_j$ of $j$-th return to $X$ with a path going along $X$ we get a sequence of closed paths defining the cycles $c_1, c_2, \dots$. These cycles represent longer and longer pieces of the geodesic. When the direction of the geodesic is uniquely ergodic, the limit

$$\lim_{N\to\infty} \frac{1}{N}\, c_N = c$$

exists and the corresponding asymptotic cycle $c \in H_1(S; \mathbb{R})$ does not depend on the starting point $x_0 \in X$. Changing the transverse interval $X$ we get a collinear asymptotic cycle.

When $S$ is a flat torus glued from a unit square, the asymptotic cycle $c$ is a vector in $H_1(\mathbb{T}^2; \mathbb{R}) = \mathbb{R}^2$ and its slope is exactly the slope of our flat geodesic in standard coordinates. When $S$ is a surface of higher genus the asymptotic cycle belongs to a $2g$-dimensional space $H_1(S; \mathbb{R}) = \mathbb{R}^{2g}$. Let us study how the cycles $c_j$ deviate from the direction of the asymptotic cycle $c$. Choose a hyperplane $W$ in $H_1(S, \mathbb{R})$ orthogonal (transversal) to the asymptotic cycle $c$ and consider a parallel projection to this screen along $c$. Projections of the cycles $c_N$ would not be necessarily bounded: directions of the cycles $c_N$ tend to direction of the asymptotic cycle $c$ provided the norms of the projections grow sublinearly with respect to $N$.

Let us observe how the projections are distributed in the screen $W$. A heuristic answer is given by Figure 3.

We see that the distribution of projections of the cycles $c_N$ in the screen $W$ is anisotropic: the projections accumulate along some line. This means that in the original space $\mathbb{R}^{2g}$ the vectors $c_N$ deviate from the asymptotic direction $L_1$ spanned by $c$ not arbitrarily but along some two-dimensional subspace $L_2$ containing $L_1$, see

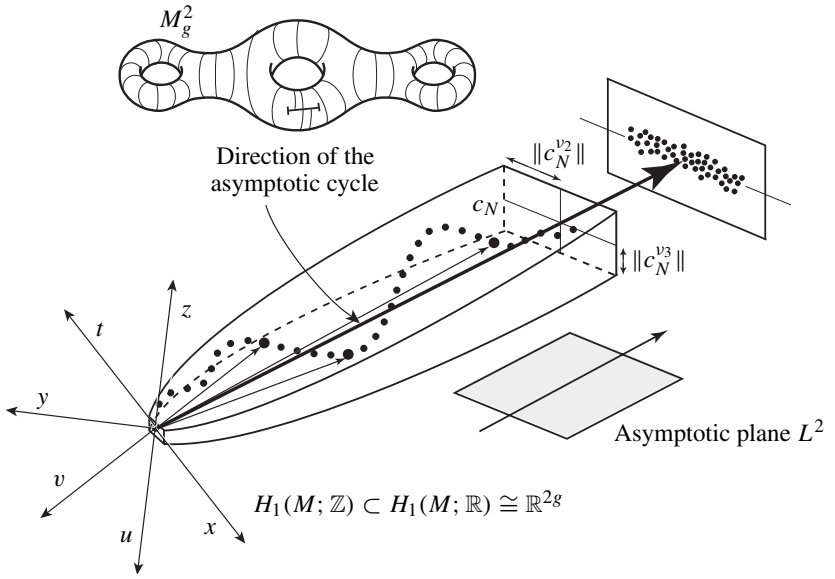



Figure 3. Deviation from the asymptotic direction exhibits anisotropic behavior: vectors deviate mainly along two-dimensional subspace, a bit more along three-dimensional subspace, etc. Their deviation from a Lagrangian $g$-dimensional subspace is already uniformly bounded.

Figure 3. Moreover, measuring the norms $\| \operatorname{proj}(c_N)\|$ of the projections we get

$$\limsup_{N\to\infty} \frac{\log \| \operatorname{proj}(c_N)\|}{\log N} = \nu_2 < 1.$$

Thus, the vector $c_N$ is located approximately in the two-dimensional plane $L_2$, and the distance from its endpoint to the line $L_1$ in $L_2$ is at most of the order $\|c_N\|^{\nu_2}$, see Figure 3.

Consider now a new screen $W_2 \perp L_2$ orthogonal to the plane $L_2$. Now the screen $W_2$ has codimension two in $H_1(S,\mathbb{R}) \simeq \mathbb{R}^{2g}$. Taking the projections of $c_N$ to $W_2$ along $L_2$ we eliminate the asymptotic directions $L_1$ and $L_2$ and we see how the vectors $c_N$ deviate from $L_2$. On the screen $W_2$ we observe the same picture as in Figure 3: the projections are again located along a one-dimensional subspace.

Coming back to the ambient space $H_1(S,\mathbb{R}) \simeq \mathbb{R}^{2g}$, this means that in the first term of approximation all vectors $c_N$ are aligned along the one-dimensional subspace $L_1$ spanned by the asymptotic cycle. In the second term of approximation, they can deviate from $L_1$, but the deviation occurs mostly in the two-dimensional subspace $L_2$, and has order $\|c_N\|^{\nu_2}$ where $\nu_2 < 1$. In the third term of approximation we see that the vectors $c_N$ may deviate from the plane $L_2$, but the deviation occurs mostly in a three-dimensional space $L_3$ and has order $\|c_N\|^{\nu_3}$ where $\nu_3 < \nu_2$.

Going on we get further terms of approximation. However, getting to a subspace $L_g$ which has half of the dimension of the ambient space we see that, in there

is no more deviation from $L_g$: the distance from any $c_N$ to $L_g$ is uniformly bounded. Note that the intersection form endows the space $H_1(S, \mathbb{R}) \simeq \mathbb{R}^{2g}$ with a natural symplectic structure. It can be checked that the resulting $g$-dimensional subspace $L_g$ is a Lagrangian subspace for this symplectic form.

A rigorous formulation of phenomena described heuristically in Figure 3 is given by the theorem below.

By convention we always consider a flat surface together with a choice of direction which is called the vertical direction, or, sometimes, "direction to the North". Using an appropriate homothethy we normalize the area of $S$ to one, so that $S \in \mathcal{H}_1(d_1, \dots, d_m)$.

We chose a point $x_0 \in S$ and a horizontal segment $X$ passing through $x_0$; by $|X|$ we denote the length of $X$. The interval $X$ is chosen in such way, that the interval exchange transformation induced by the vertical flow has the minimal possible number $n = 2g + m - 1$ of subintervals under exchange. (Actually, almost any other choice of $X$ would also work.) We consider a geodesic ray $\gamma$ emitted from $x_0$ in the vertical direction. (If $x_0$ is a saddle point, there are several outgoing vertical geodesic rays; choose any of them.) Each time when $\gamma$ intersects $X$ we join the point $x_N$ of intersection and the starting point $x_0$ along $X$ producing a closed path. We denote the homology class of the corresponding loop by $c_N$.

Let $\omega$ be the holomorphic 1-form representing $S$; let $g$ be genus of $S$. Choose some Euclidean metric in $H_1(S; \mathbb{R}) \simeq \mathbb{R}^{2g}$ which would allow to measure a distance from a vector to a subspace. Let by convention $\log(0) = -\infty$.

**Theorem 1.** *For almost any flat surface $S$ in any stratum $\mathcal{H}_1(d_1, \dots, d_m)$ there exists a flag of subspaces*

$$L_1 \subset L_2 \subset \cdots \subset L_g \subset H_1(S; \mathbb{R})$$

*in the first homology group of the surface with the following properties.*

*Choose any starting point $x_0 \in X$ in the horizontal segment $X$. Consider the corresponding sequence $c_1, c_2, \dots$ of cycles.*

*– The following limit exists:*

$$|X| \lim_{N\to\infty} \frac{1}{N} c_N = c,$$

*where the nonzero asymptotic cycle $c \in H_1(M_g^2; \mathbb{R})$ is Poincaré dual to the cohomology class of $\omega_0 = \mathrm{Re}[\omega]$, and the one-dimensional subspace $L_1 = \langle c \rangle_{\mathbb{R}}$ is spanned by $c$.*

*– For any $j = 1, \dots, g-1$ one has*

$$\limsup_{N\to\infty} \frac{\log \mathrm{dist}(c_N, L_j)}{\log N} = \nu_{j+1}$$

*and*

$$\mathrm{dist}(c_N, L_g) \leq \mathrm{const},$$

*where the constant depends only on S and on the choice of the Euclidean structure in the homology space.*

*The numbers* $2, 1 + \nu_2, \ldots, 1 + \nu_g$ *are the top g Lyapunov exponents of the Teichmüller geodesic flow on the corresponding connected component of the stratum* $\mathcal{H}(d_1, \ldots, d_m)$*; in particular, they do not depend on the individual generic flat surface S in the connected component.*

It should be stressed, that the theorem above was formulated in [Z3] as a conditional statement: under the conjecture that $\nu_g > 0$ there exist a Lagrangian subspace $L_g$ such that the cycles are in a bounded distance from $L_g$; under the further conjecture that all the exponents $\nu_j$, for $j = 2, \ldots, g$, are distinct, there is a complete Lagrangian flag (i.e. the dimensions of the subspaces $L_j$, where $j = 1, 2, \ldots, g$, rise each time by one). These two conjectures were later proved by G. Forni [Fo1] and by A. Avila and M. Viana [AvVi] correspondingly.

Currently there are no methods of calculation of individual Lyapunov exponents $\nu_j$ (though there is some experimental knowledge of their approximate values). Nevertheless, for any connected component of any stratum (and, more generally, for any $\mathrm{GL}^+(2;\mathbb{R})$-invariant suborbifold) it is possible to evaluate the *sum* of the Lyapunov exponents $\nu_1 + \cdots + \nu_g$, where $g$ is the genus. The formula for this sum was discovered by M. Kontsevich; morally, it is given in terms of characteristic numbers of some natural vector bundles over the strata $\mathcal{H}(d_1, \ldots, d_m)$, see [K]. Another interpretation of this formula was found by G. Forni [Fo1]; see also a very nice formalization of these results in the survey of R. Krikorian [Kr]. For some special $\mathrm{GL}^+(2;\mathbb{R})$-invariant suborbifolds the corresponding vector bundles might have equivariant subbundles, which provides additional information on corresponding subcollections of the Lyapunov exponents, or even gives their explicit values in some cases, like in the case of Teichmüller curves considered in the paper of I. Bouw and M. Möller [BMö].

Theorem 1 illustrates a phenomenon of *deviation spectrum*. It was proved by G. Forni in [Fo1] that ergodic sums of smooth functions on an interval along trajectories of interval exchange transformations, and ergodic integrals of smooth functions on flat surfaces along trajectories of directional flows have deviation spectrum analogous to the one described in Theorem 1. L. Flaminio and G. Forni showed that the same phenomenon can be observed for other parabolic dynamical systems, for example, for the horocycle flow on compact surfaces of constant negative curvature [FlFo].

**Idea of the proof: renormalization.** The reason why the deviation of the cycles $c_j$ from the asymptotic direction is governed by the Teichmüller geodesic flow is illustrated in Figure 4. In a sense, we follow the initial ideas of H. Masur [M1] and of W. Veech [V1].

Fix a horizontal segment $X$ and emit a vertical trajectory from some point $x$ in $X$. When the trajectory intersects $X$ for the first time join the corresponding point $T(x)$ to the original point $x$ along $X$ to obtain a closed loop. Here $T\colon X \to X$

denotes the first return map to the transversal $X$ induced by the vertical flow. Denote by $c(x)$ the corresponding cycle in $H_1(S;\mathbb{Z})$. Let the interval exchange transformation $T\colon X \to X$ decompose $X$ into $n$ subintervals $X_1 \sqcup \cdots \sqcup X_n$. It is easy to see that the "first return cycle" $c(x)$ is piecewise constant: we have $c(x) = c(x') =: c(X_j)$ whenever $x$ and $x'$ belong to the same subinterval $X_j$, see Figure 4. It is easy to see that

$$c_N(x) = c(x) + c(T(x)) + \cdots + c(T^{N-1}(x)).$$

The average of this sum with respect to the "time" $N$ tends to the asymptotic cycle $c$. We need to study the deviation of this sum from the value $N \cdot c$. To do this consider a shorter subinterval $X'$ as in Figure 4. Its length is chosen in such way, that the first return map of the vertical flow again induces an interval exchange transformation $T'\colon X' \to X'$ of $n$ subintervals. New first return cycles $c'(X'_k)$ to the interval $X'$ are expressed in terms of the initial first return cycles $c(X_j)$ by the linear relations below; the lengths $|X'_k|$ of subintervals of the new partition $X' = X'_1 \sqcup \cdots \sqcup X'_m$ are expressed in terms of the lengths $|X_j|$ of subintervals of the initial partition by dual linear relations:

$$c'(X'_k) = \sum_{j=1}^{n} A_{jk} \cdot c(X_j), \quad |X_j| = \sum_{k=1}^{n} A_{jk} \cdot |X'_k|,$$

where a nonnegative integer matrix $A_{jk}$ is completely determined by the initial interval exchange transformation $T\colon X \to X$ and by the choice of $X' \subset X$.

To construct the cycle $c_N$ representing a long piece of leaf of the vertical foliation we followed the trajectory $x, T(x), \ldots, T^{N-1}(x)$ of the initial interval exchange transformation $T\colon X \to X$ and computed the corresponding ergodic sum. Passing to a shorter horizontal interval $X' \subset X$ we can follow the trajectory $x, T'(x), \ldots, (T')^{N'-1}(x)$ of the new interval exchange transformation $T'\colon X' \to X'$ (provided $x \in X'$). Since the subinterval $X'$ is shorter than $X$ we cover the initial piece of trajectory of the vertical flow in a smaller number $N'$ of steps. In other words, passing from $T$ to $T'$ we accelerate the time: it is easy to see that the trajectory $x, T'(x), \ldots, (T')^{N'-1}(x)$ follows the trajectory $x, T(x), \ldots, T^{N-1}(x)$ but jumps over several iterations of $T$ at a time.

This approach would not be efficient if the new first return map $T'\colon X' \to X'$ would be more complicated than the initial one. But we know that passing from $T$ to $T'$ we stay within a family of interval exchange transformations of some fixed number $n$ of subintervals, and, moreover, that the new "first return cycles" $c'(X'_k)$ and the lengths $|X'_k|$ of the new subintervals are expressed in terms of the initial ones by means of the $n \times n$-matrix $A$, which depends only on the choice of $X' \subset X$ and which can be easily computed.

Our strategy can be now formulated as follows. One can define an explicit algorithm (generalizing Euclidean algorithm) which canonically associates to an interval exchange transformation $T\colon X \to X$ some specific subinterval $X' \subset X$ and, hence,

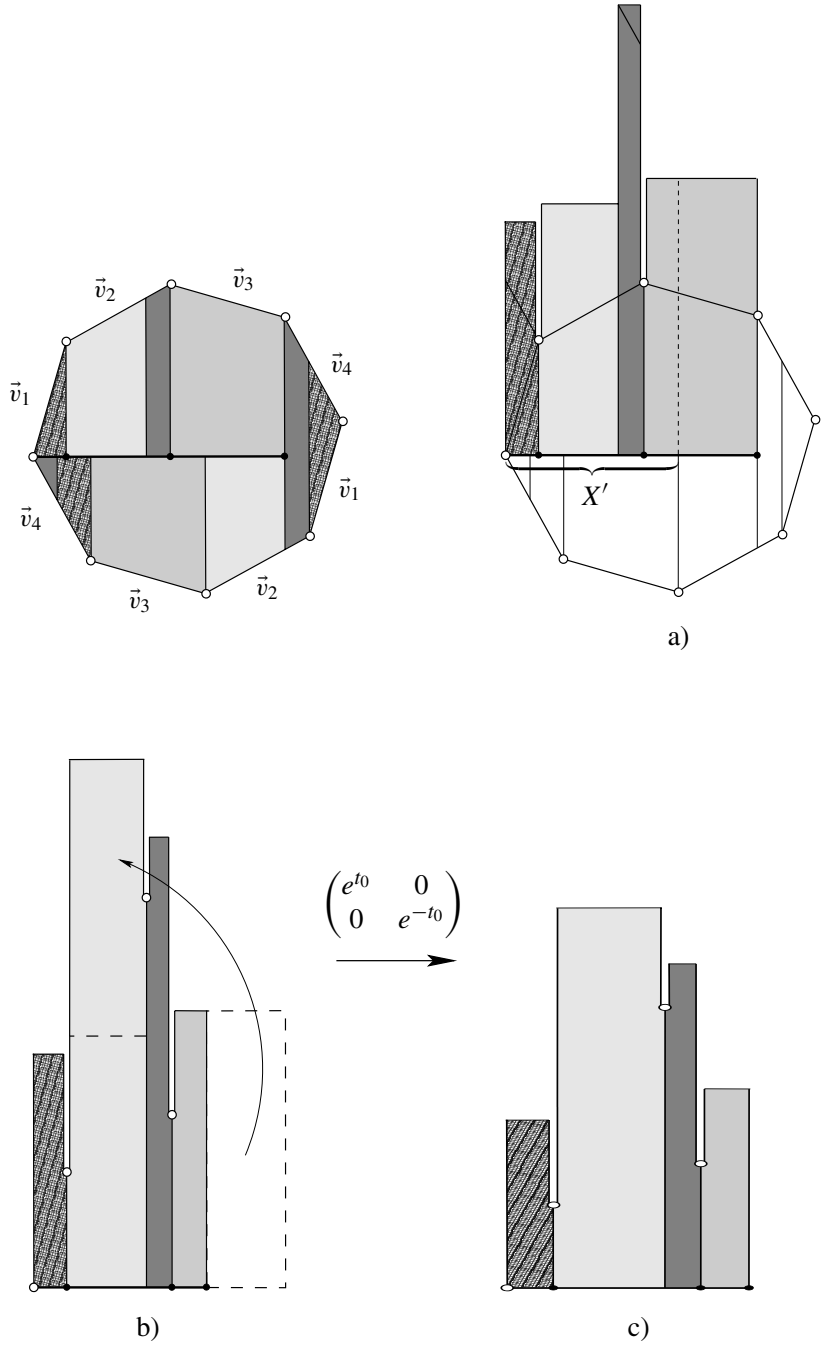



Figure 4. Idea of renormalization. a) Unwrap the flat surface into "zippered rectangles". b) Shorten the base of the corresponding zippered rectangles. c) Expand the resulting tall and narrow zippered rectangle horizontally and contract it vertically by same factor $e^{t_0}$.

a new interval exchange transformation $T'\colon X' \to X'$. Similarly to the Euclidean algorithm our algorithm is invariant under proportional rescaling of $X$ and $X'$, so, when we find it convenient, we can always rescale the length of the interval to one. This algorithm can be considered as a map $\mathcal{T}$ from the space of all interval exchange transformations of a given number $n$ of subintervals to itself. Applying recursively this algorithm we construct a sequence of subintervals $X = X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \cdots$ and a sequence of matrices $A = A(T^{(0)}), A(T^{(1)}), \ldots$ describing transitions form interval exchange transformation $T^{(r)}\colon X^{(r)} \to X^{(r)}$ to interval exchange transformation $T^{(r+1)}\colon X^{(r+1)} \to X^{(r+1)}$. Taking a product $A^{(s)} = A(T^{(0)}) \cdot A(T^{(1)}) \ldots A(T^{(s-1)})$ we can immediately express the "first return cycles" to a microscopic subinterval $X^{(s)}$ in terms of the initial "first return cycles" to $X$. Considering now the matrices $A$ as the values of a matrix-valued function on the space of interval exchange transformations, we realize that we study the products of matrices $A$ along the orbits $T^{(0)}, T^{(1)}, \ldots, T^{(s-1)}$ of the map on the space of interval exchange transformations. When the map is ergodic with respect to a finite measure, the properties of these products are described by the Oseledets theorem, and the cycles $c_N$ have a deviation spectrum governed by the Lyapunov exponents of the cocycle $A$ on the space of interval exchange transformations.

Note that the first return cycle to the subinterval $X^{(s)}$ (which is very short) represents the cycle $c_N$ corresponding to a very long trajectory $x, T(x), \ldots, T^{N-1}(x)$ of the initial interval exchange transformation. In other words, our renormalization procedure $\mathcal{T}$ plays a role of a time acceleration machine: morally, instead of getting the cycle $c_N$ by following a trajectory $x, T(x), \ldots, T^{N-1}(x)$ of the initial interval exchange transformation for the exponential time $N \sim \exp(\text{const} \cdot s)$ we obtain the cycle $c_N$ applying only $s$ steps of the renormalization map $\mathcal{T}$ on the space of interval exchange transformations.

It remains to establish the relation between the cocycle $A$ over the map $\mathcal{T}$ and the Teichmüller geodesic flow. Conceptually, this relation was elaborated in the original paper of W. Veech [V1].

First let us discuss how can one "almost canonically" (that is up to a finite ambiguity) choose a zippered rectangles representation of a flat surface. Note that Figure 4 suggests the way which allows to obtain infinitely many zippered rectangles representations of the same flat surface: we chop an appropriate rectangle on the right, put it atop the corresponding rectangle and then repeat the procedure recursively. This resembles the situation with a representation of a flat torus by a parallelogram: a point of the fundamental domain in Figure 2 provides a canonical representative though any point of the corresponding $\mathrm{SL}(2, \mathbb{Z})$-orbit represents the same flat torus. A "canonical" zippered rectangles decomposition of a flat surface also belongs to some fundamental domain. Following W. Veech one can define the fundamental domain in terms of some specific choice of a "canonical" horizontal interval $X$. Namely, let us position the left endpoint of $X$ at a conical singularity. Let us choose the length of $X$ in such way that the interval exchange transformation $T\colon X \to X$ induced by the first return of the vertical flow to $X$ has minimal possible number $n = 2g + m - 1$

of subintervals under exchange. Among all such horizontal segments $X$ choose the shortest one, which length is greater than or equal to one. This construction is applicable to almost all flat surfaces; the finite ambiguity corresponds to the finite freedom in the choice of the conical singularity and in the choice of the horizontal ray adjacent to it.

Since the interval $X$ defines a decomposition of (almost any) flat surface into "zippered rectangles" (see Figure 4) we can pass from the space of flat surfaces to the space of zippered rectangles (which can be considered as a finite ramified covering over the space of flat surfaces). Teichmüller geodesic flow lifts naturally to the space of zippered rectangles. It acts on zippered rectangles by expansion in horizontal direction and contraction in vertical direction; i.e. the zippered rectangles are modified by linear transformations $\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$. However, as soon as the Teichmüller geodesic flow brings us out of the fundamental domain, we have to modify the zippered rectangles decomposition to the "canonical one" corresponding to the fundamental domain. (Compare to Figure 2 where the Teichmüller geodesic flow corresponds to the standard geodesic flow in the hyperbolic metric on the upper half-plane.) The corresponding modification of zippered rectangles (chop an appropriate rectangle on the right, put it atop the corresponding rectangle; repeat the procedure several times, if necessary) is illustrated in Figure 4.

Now everything is ready to establish the relation between the Teichmüller geodesic flow and the map $\mathcal{T}$ on the space of interval exchange transformations.

Consider some codimension one subspace $\Upsilon$ in the space of zippered rectangles transversal to the Teichmüller geodesic flow. Say, $\Upsilon$ might be defined by the requirement that the base $X$ of the zippered rectangles decomposition has length one, $|X| = 1$. This is the choice in the original paper of W. Veech [V1]; under this choice $\Upsilon$ represents part of the boundary of the fundamental domain in the space of zippered rectangles. Teichmüller geodesic flow defines the first return map $\mathcal{S}\colon \Upsilon \to \Upsilon$ to the section $\Upsilon$. The map $\mathcal{S}$ can be described as follows. Take a flat surface of unit area decomposed into zippered rectangles $Z$ with the base $X$ of length one. Apply expansion in horizontal direction and contraction in vertical direction. For some $t_0(Z)$ the deformed zippered rectangles can be rearranged as in Figure 4 to get back to the base of length one; the result is the image of the map $\mathcal{S}$. Actually, we can first apply the rearrangement as in Figure 4 to the initial zippered rectangles $Z$ and then apply the transformation $\begin{pmatrix} e^{t_0} & 0 \\ 0 & e^{-t_0} \end{pmatrix}$ – the two operations commute. This gives, in particular, an explicit formula for $t_0(Z)$. Namely let $|X_n|$ be the width of the rightmost rectangle and let $|X_k|$ be the width of the rectangle, which top horizontal side is glued to the rightmost position at the base $X$. (For the upper zippered rectangle decomposition in Figure 4 we have $n = 4$ and $k = 2$.) Then

$$t_0 = -\log\bigl(1 - \min(|X_n|, |X_k|)\bigr).$$

Recall that a decomposition of a flat surface into zippered rectangles naturally defines an interval exchange transformation – the first return map of the vertical flow

to the base $X$ of zippered rectangles. Hence, the map $\mathcal{S}$ of the subspace $\Upsilon$ of zippered rectangles defines an induced map on the space of interval exchange transformations. It remains to note that this induced map is exactly the map $\mathcal{T}$. In other words, the map $\mathcal{S}\colon \Upsilon \to \Upsilon$ induced by the first return of the Teichmüller geodesic flow to the subspace $\Upsilon$ of zippered rectangles is the suspension of the map $\mathcal{T}$ on the space of interval exchange transformations.

We complete with a remark concerning the choice of a section. The natural section $\Upsilon$ chosen in the original paper of W. Veech [V1] is in a sense too large: the corresponding invariant measure (induced from the measure on the space of flat surfaces) is infinite. Choosing an appropriate subset $\Upsilon' \subset \Upsilon$ one can get finite invariant measure. Moreover, the subset $\Upsilon'$ can be chosen in such way that the corresponding first return map $\mathcal{S}'\colon \Upsilon' \to \Upsilon'$ of the Teichmüller geodesic flow is a suspension of some natural map $\mathcal{G}$ on the space of interval exchange transformations, see [Z1]. According to the results of H. Masur [M1] and W. Veech [V1] the Teichmüller geodesic flow is ergodic which implies ergodicity of the maps $\mathcal{S}'$ and $\mathcal{G}$. To apply Oseledets theorem one should, actually, consider the induced cocycle $B$ over this new map $\mathcal{G}$ instead of the cocycle $A$ over the map $\mathcal{T}$ described above.

## 2. Closed geodesics on flat surfaces

Consider a flat surface $S$; we always assume that the flat metric on $S$ has trivial holonomy, and that the surface $S$ has finite number of cone-type singularities. By convention a flat surface is endowed with a choice of direction, refereed to as a "vertical direction", or as a "direction to the North". Since the flat metric has trivial holonomy, this direction can be transported in a unique way to any point of the surface.

A geodesic segment joining two conical singularities and having no conical points in its interior is called *saddle connection*. The case when boundaries of a saddle connection coincide is not excluded: a saddle connection might join a conical point to itself. In this section we study saddle connections and closed regular geodesics on a generic flat surface $S$ of genus $g \geq 2$. In particular, we count them and we explain the following curious phenomenon: saddle connections and closed regular geodesics often appear in pairs, triples, etc of parallel saddle connections (correspondingly closed regular geodesics) of the same direction and length. When all saddle connections (closed regular geodesics) in such configuration are short the corresponding flat surface is almost degenerate; it is located close to the boundary of the moduli space. A description of possible configurations of parallel saddle connections (closed geodesics) gives us a description of the multidimensional "cusps" of the strata.

The results of this section are based on the joint work with A. Eskin and H. Masur [EMZ] and on their work [MZ]. A series of beautiful results developing the counting problems considered here were recently obtained by Ya. Vorobets [Vo].

**Counting closed geodesics and saddle connections.** Closed geodesics on flat surfaces of higher genera have some similarities with ones on the torus. Suppose that we have a regular closed geodesic passing through a point $x_0 \in S$. Emitting a geodesic from a nearby point $x$ in the same direction we obtain a parallel closed geodesic of the same length as the initial one. Thus, closed geodesics appear in families of parallel closed geodesics. However, in the torus case every such family fills the entire torus while a family of parallel regular closed geodesics on a flat surfaces of higher genus fills only part of the surface. Namely, it fills a flat cylinder having a conical singularity on each of its boundaries. Typically, a maximal cylinder of closed regular geodesics is bounded by a pair of closed saddle connections. Reciprocally, any saddle connection joining a conical point $P$ to itself and coming back to $P$ at the angle $\pi$ bounds a cylinder filled with closed regular geodesics.

A geodesic representative of a homotopy class of a curve on a flat surface is realized in general by a broken line of geodesic segments with vertices at conical points. By convention we consider only closed *regular* geodesics (which by definition do not pass through conical points) or saddle connections (which by definition do not have conical points in its interior). Everywhere in this section we normalize the area of a flat surface to one.

Let $N_{\mathrm{sc}}(S, L)$ be the number of saddle connections of length at most $L$ on a flat surface $S$. Let $N_{\mathrm{cg}}(S, L)$ be the number of maximal cylinders filled with closed regular geodesics of length at most $L$ on $S$. It was proved by H. Masur that for any flat surface $S$ both counting functions $N(S, L)$ grow quadratically in $L$. Namely, there exist constants $0 < \mathrm{const}_1(S) < \mathrm{const}_2(S) < \infty$ such that

$$\mathrm{const}_1(S) \leq N(S, L)/L^2 \leq \mathrm{const}_2(S)$$

for $L$ sufficiently large. Recently Ya. Vorobets has obtained uniform estimates for the constants $\mathrm{const}_1(S)$ and $\mathrm{const}_2(S)$ which depend only on the genus of $S$, see [Vo]. Passing from *all* flat surfaces to *almost all* surfaces in a given connected component of a given stratum one gets a much more precise result, see [EM]:

**Theorem** (A. Eskin and H. Masur). *For almost all flat surfaces $S$ in any stratum $\mathcal{H}(d_1, \ldots, d_m)$ the counting functions $N_{\mathrm{sc}}(S, L)$ and $N_{\mathrm{cg}}(S, L)$ have exact quadratic asymptotics*

$$\lim_{L\to\infty} \frac{N_{\mathrm{sc}}(S, L)}{\pi L^2} = c_{\mathrm{sc}}(S), \quad \lim_{L\to\infty} \frac{N_{\mathrm{cg}}(S, L)}{\pi L^2} = c_{\mathrm{cg}}(S).$$

*Moreover, the Siegel–Veech constants $c_{\mathrm{sc}}(S)$ (correspondingly $c_{\mathrm{cg}}(S)$) coincide for almost all flat surfaces $S$ in each connected component $\mathcal{H}_1^{\mathrm{comp}}(d_1, \ldots, d_m)$ of the stratum.*

**Phenomenon of higher multiplicities.** Note that the direction to the North is well-defined even at a conical point of a flat surface, moreover, at a conical point $P_1$ with

a cone angle $2\pi k$ we have $k$ different directions to the North! Consider some saddle connection $\gamma_1 = [P_1 P_2]$ with an endpoint at $P_1$. Memorize its direction, say, let it be the North–West direction. Let us launch a geodesic from the same starting point $P_1$ in one of the remaining $k-1$ North–West directions. Let us study how big is the chance to hit $P_2$ ones again, and how big is the chance to hit it after passing the same distance as before. We do not exclude the case $P_1 = P_2$. Intuitively it is clear that the answer to the first question is: "the chances are low" and to the second one is "the chances are even lower". This makes the following theorem (see [EMZ]) somehow counterintuitive:

**Theorem 2** (A. Eskin, H. Masur, A. Zorich). *For almost any flat surface $S$ in any stratum and for any pair $P_1$, $P_2$ of conical singularities on $S$ the function $N_2(S, L)$ counting the number of pairs of parallel saddle connections of the same length joining $P_1$ to $P_2$ has exact quadratic asymptotics*

$$\lim_{L\to\infty} \frac{N_2(S, L)}{\pi L^2} = c_2 > 0,$$

*where the Siegel–Veech constant $c_2$ depends only on the connected component of the stratum and on the cone angles at $P_1$ and $P_2$.*

*For almost all flat surfaces $S$ in any stratum one cannot find neither a single pair of parallel saddle connections on $S$ of different length, nor a single pair of parallel saddle connections joining different pairs of singularities.*

Analogous statements (with some reservations for specific connected components of certain strata) can be formulated for arrangements of $3, 4, \dots$ parallel saddle connections. The situation with closed regular geodesics is similar: they might appear (also with some exceptions for specific connected components of certain strata) in families of $2, 3, \dots, g-1$ distinct maximal cylinders filled with parallel closed regular geodesics of equal length. A general formula for the Siegel–Veech constant in the corresponding quadratic asymptotics is presented at the end of this section, while here we want to discuss the numerical values of Siegel–Veech constants in a simple concrete example. We consider the principal strata $\mathcal{H}(1, \dots, 1)$ in small genera. Let $N_{k\_\mathrm{cyl}}(S, L)$ be the corresponding counting function, where $k$ is the number of distinct maximal cylinders filled with parallel closed regular geodesics of equal length bounded by $L$. Let

$$c_{k\_\mathrm{cyl}} = \lim_{L\to\infty} \frac{N_{k\_\mathrm{cyl}}(S, L)}{\pi L^2}.$$

The table below (extracted from [EMZ]) presents the values of $c_{k\_\mathrm{cyl}}$ for $g = 1, \dots, 4$. Note that for a generic flat surface $S$ of genus $g$ a configuration of $k \geq g$ cylinders is not realizable, so we do not fill the corresponding entry.

| $k$ | $g=1$ | $g=2$ | $g=3$ | $g=4$ |
|---|---|---|---|---|
| 1 | $\frac{1}{2}\cdot\frac{1}{\zeta(2)}\approx 0.304$ | $\frac{5}{2}\cdot\frac{1}{\zeta(2)}\approx 1.52$ | $\frac{36}{7}\cdot\frac{1}{\zeta(2)}\approx 3.13$ | $\frac{3150}{377}\cdot\frac{1}{\zeta(2)}\approx 5.08$ |
| 2 | – | – | $\frac{3}{14}\cdot\frac{1}{\zeta(2)}\approx 0.13$ | $\frac{90}{377}\cdot\frac{1}{\zeta(2)}\approx 0.145$ |
| 3 | – | – | – | $\frac{5}{754}\cdot\frac{1}{\zeta(2)}\approx 0.00403$ |

Comparing these values we see, that our intuition was not quite misleading. Morally, in genus $g = 4$ a closed regular geodesic belongs to a one-cylinder family with "probability" 97.1%, to a two-cylinder family with "probability" 2.8% and to a three-cylinder family with "probability" only 0.1% (where "probabilities" are calculated proportionally to the Siegel–Veech constants $5.08 : 0.145 : 0.00403$).

**Rigid configurations of saddle connections and "cusps" of the strata.** A saddle connection or a regular closed geodesic on a flat surface $S$ persists under small deformations of $S$ inside the corresponding stratum. It might happen that any deformation of a given flat surface which shortens some specific saddle connection necessarily shortens some other saddle connections. We say that a collection $\{\gamma_1, \dots, \gamma_n\}$ of saddle connections is *rigid* if any sufficiently small deformation of the flat surface inside the stratum preserves the proportions $|\gamma_1| : |\gamma_2| : \cdots : |\gamma_n|$ of the lengths of all saddle connections in the collection. It was shown in [EMZ] that all saddle connections in any rigid collection are *homologous*. Since their directions and lengths can be expressed in terms of integrals of the holomorphic 1-form $\omega$ along corresponding paths, this implies that homologous saddle connections $\gamma_1, \dots, \gamma_n$ are parallel and have equal length and either all of them join the same pair of distinct singular points, or all $\gamma_i$ are closed loops.

This implies that when saddle connections in a rigid collection are contracted by a continuous deformation, the limiting flat surface generically decomposes into several connected components represented by nondegenerate flat surfaces $S'_1, \dots, S'_k$, see Figure 5, where $k$ might vary from one to the genus of the initial surface. Let the initial surface $S$ belong to a stratum $\mathcal{H}(d_1, \dots, d_m)$. Denote the set with multiplicities $\{d_1, \dots, d_m\}$ by $\beta$. Let $\mathcal{H}(\beta'_j)$ be the stratum ambient for $S'_j$. The stratum $\mathcal{H}(\beta') = \mathcal{H}(\beta'_1) \sqcup \cdots \sqcup \mathcal{H}(\beta'_k)$ of disconnected flat surfaces $S'_1 \sqcup \cdots \sqcup S'_k$ is referred to as a *principal boundary* stratum of the stratum $\mathcal{H}(\beta)$. For any connected component of any stratum $\mathcal{H}(\beta)$ the paper [EMZ] describes all principal boundary strata; their

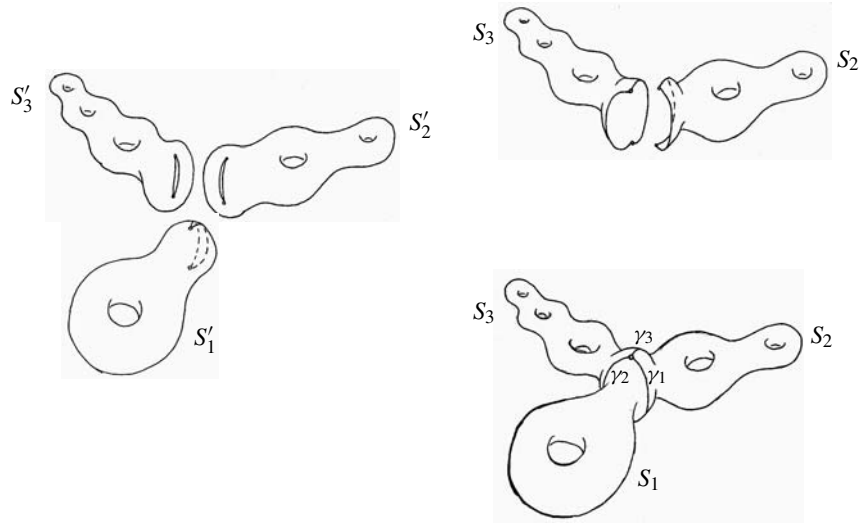



Figure 5. Multiple homologous saddle connections, topological picture (after [EMZ]).

union is called the principal boundary of the corresponding connected component of $\mathcal{H}(\beta)$.

The paper [EMZ] also presents the inverse construction. Consider any flat surface $S'_1 \sqcup \cdots \sqcup S'_k \in \mathcal{H}(\beta')$ in the principal boundary of $\mathcal{H}(\beta)$; consider a sufficiently small value of a complex parameter $\varepsilon \in \mathbb{C}$. One can reconstruct the flat surface $S \in \mathcal{H}(\beta)$ endowed with a collection of homologous saddle connections $\gamma_1, \ldots, \gamma_n$ such that $\int_{\gamma_i} \omega = \varepsilon$, and such that degeneration of $S$ contracting the saddle connections $\gamma_i$ in the collection gives the surface $S'_1 \sqcup \cdots \sqcup S'_k$. This inverse construction involves several surgeries of the flat structure. Having a disconnected flat surface $S'_1 \sqcup \cdots \sqcup S'_k$ one applies an appropriate surgery to each $S'_j$ producing a surface $S_j$ with boundary. The surgery depends on the parameter $\varepsilon$: the boundary of each $S_j$ is composed from two geodesic segments of lengths $|\varepsilon|$; moreover, the boundary components of $S_j$ and $S_{j+1}$ are compatible, which allows to glue the compound surface $S$ from the collection of surfaces with boundary, see Figure 5 as an example.

A collection $\gamma = \{\gamma_1, \ldots, \gamma_n\}$ of homologous saddle connections determines the following data on combinatorial geometry of the decomposition $S \setminus \gamma$: the number of components, their boundary structure, the singularity data for each component, the cyclic order in which the components are glued to each other. These data are referred to as a *configuration* of homologous saddle connections. A configuration $\mathcal{C}$ uniquely determines the corresponding boundary stratum $\mathcal{H}(\beta'_{\mathcal{C}})$; it does not depend on the collection $\gamma$ of homologous saddle connections representing the configuration $\mathcal{C}$.

The constructions above explain how configurations $\mathcal{C}$ of homologous saddle connections on flat surfaces $S \in \mathcal{H}(\beta)$ determine the "cusps" of the stratum $\mathcal{H}(\beta)$. Consider a subset $\mathcal{H}^{\varepsilon}_1(\beta) \subset \mathcal{H}(\beta)$ of surfaces of area one having a saddle connec-

tion shorter than $\varepsilon$. Up to a subset $\mathcal{H}_1^{\varepsilon,\text{thin}}(\beta)$ of negligibly small measure the set $\mathcal{H}_1^{\varepsilon,\text{thick}}(\beta) = \mathcal{H}_1^{\varepsilon}(\beta) \setminus \mathcal{H}_1^{\varepsilon,\text{thin}}(\beta)$ might be represented as a disjoint union

$$\mathcal{H}_1^{\varepsilon,\text{thick}}(\beta) \approx \bigsqcup_{\mathcal{C}} \mathcal{H}_1^{\varepsilon}(\mathcal{C})$$

of neighborhoods $\mathcal{H}_1^{\varepsilon}(\mathcal{C})$ of the corresponding "cusps" $\mathcal{C}$. Here $\mathcal{C}$ runs over a finite set of configurations admissible for the given stratum $\mathcal{H}_1(\beta)$; this set is explicitly described in [EMZ].

When a configuration $\mathcal{C}$ is composed from homologous saddle connections joining *distinct* zeroes, the neighborhood $\mathcal{H}_1^{\varepsilon}(\mathcal{C})$ of the cusp $\mathcal{C}$ has the structure of a fiber bundle over the corresponding boundary stratum $\mathcal{H}(\beta'_{\mathcal{C}})$ (up to a difference in a set of a negligibly small measure). A fiber of this bundle is represented by a finite cover over the Euclidean disc of radius $\varepsilon$ ramified at the center of the disc. Moreover, the canonical measure in $\mathcal{H}_1^{\varepsilon}(\mathcal{C})$ decomposes into a product measure of the canonical measure in the boundary stratum $\mathcal{H}(\beta'_{\mathcal{C}})$ and the Euclidean measure in the fiber (see [EMZ]), so

$$\text{Vol}\left(\mathcal{H}_1^{\varepsilon}(\mathcal{C})\right) = (\text{combinatorial factor}) \cdot \pi\varepsilon^2 \cdot \prod_{j=1}^{k} \text{Vol}\,\mathcal{H}_1(\beta'_j) \;+\; o(\varepsilon^2). \qquad (1)$$

**Remark.** We warn the reader that the correspondence between compactification of the moduli space of Abelian differentials and the Deligne–Mumford compactification of the underlying moduli space of curves is not straightforward. In particular, the desingularized stable curve corresponding to the limiting flat surface generically *is not* represented as a union of Riemann surfaces corresponding to $S'_1, \ldots, S'_k$ – the stable curve might contain more components.

**Evaluation of the Siegel–Veech constants.** Consider a flat surface $S$. To every closed regular geodesic $\gamma$ on $S$ we can associate a vector $\vec{v}(\gamma)$ in $\mathbb{R}^2$ having the length and the direction of $\gamma$. In other words, $\vec{v} = \int_\gamma \omega$, where we consider a complex number as a vector in $\mathbb{R}^2 \simeq \mathbb{C}$. Applying this construction to all closed regular geodesic on $S$ we construct a discrete set $V(S) \subset \mathbb{R}^2$. Consider the following operator $f \mapsto \hat{f}$ from functions with compact support on $\mathbb{R}^2$ to functions on a connected component $\mathcal{H}_1^{\text{comp}}(\beta)$ of the stratum $\mathcal{H}_1(\beta) = \mathcal{H}_1(d_1, \ldots, d_m)$:

$$\hat{f}(S) := \sum_{\vec{v} \in V(S)} f(\vec{v}).$$

Function $\hat{f}(S)$ generalizes the counting function $N_{\text{cg}}(S, L)$ introduced in the beginning of this section. Namely, when $f = \chi_L$ is the characteristic function $\chi_L$ of the disc of radius $L$ with the center at the origin of $\mathbb{R}^2$, the function $\hat{\chi}_L(S)$ counts the number of regular closed geodesics of length at most $L$ on a flat surface $S$.

**Theorem** (W. Veech). *For any function $f\colon \mathbb{R}^2 \to \mathbb{R}$ with compact support the following equality is valid:*

$$\frac{1}{\operatorname{Vol} \mathcal{H}_1^{\mathrm{comp}}(\beta)} \int_{\mathcal{H}_1^{\mathrm{comp}}(\beta)} \hat{f}(S)\, dv_1 = C \int_{\mathbb{R}^2} f(x, y)\, dx\, dy, \tag{2}$$

*where the constant $C$ does not depend on the function $f$.*

Note that this is an exact equality. In particular, choosing the characteristic function $\chi_L$ of a disc of radius $L$ as a function $f$ we see that for any positive $L$ the *average* number of closed regular geodesics not longer than $L$ on flat surfaces $S \in \mathcal{H}_1^{\mathrm{comp}}(\beta)$ is exactly $C \cdot \pi L^2$, where the Siegel–Veech constant $C$ does not depend on $L$, but only on the connected component $\mathcal{H}_1^{\mathrm{comp}}(\beta)$.

The theorem of Eskin and Masur cited above tells that for large values of $L$ one gets approximate equality $\hat{\chi}_L(S) \approx c_{\mathrm{cg}} \cdot \pi L^2$ "pointwisely" for almost all individual flat surfaces $S \in \mathcal{H}_1^{\mathrm{comp}}(d_1, \dots, d_m)$. It is proved in [EM] that the corresponding Siegel–Veech constant $c_{\mathrm{cg}}$ coincides with the constant $C$ in equation (2) above.

Actually, the same technique can be applied to count separately pairs, triples, or any other specific configurations $\mathcal{C}$ of homologous saddle connections. Every time when we find a collection of homologous saddle connections $\gamma_1, \dots, \gamma_n$ representing the chosen configuration $\mathcal{C}$ we construct a vector $\vec{v} = \int_{\gamma_i} \omega$. Since all $\gamma_1, \dots, \gamma_n$ are homologous, we can take any of them as $\gamma_i$. Taking all possible collections of homologous saddle connections on $S$ representing the fixed configuration $\mathcal{C}$ we construct new discrete set $V_{\mathcal{C}}(S) \subset \mathbb{R}^2$ and new functional $f \mapsto \hat{f}_{\mathcal{C}}$. Theorem of Eskin and Masur and theorem of Veech [V4] presented above are valid for $\hat{f}_{\mathcal{C}}$. The corresponding Siegel–Veech constant $c(\mathcal{C})$ responsible for the quadratic growth rate $N_{\mathcal{C}}(S, L) \sim c(\mathcal{C}) \cdot \pi L^2$ of the number of collections of homologous saddle connections of the type $\mathcal{C}$ on an individual generic flat surface $S$ coincides with the constant $C(\mathcal{C})$ in the expression analogous to (2).

Formula (2) can be applied to $\hat{\chi}_L$ for any value of $L$. In particular, instead of taking large $L$ we can choose a very small $L = \varepsilon \ll 1$. The corresponding function $\hat{\chi}_\varepsilon(S)$ counts how many collections of parallel $\varepsilon$-short saddle connections (closed geodesics) of the type $\mathcal{C}$ we can find on a flat surface $S \in \mathcal{H}_1^{\mathrm{comp}}(\beta)$. For the flat surfaces $S$ outside of the subset $\mathcal{H}_1^{\varepsilon}(\mathcal{C}) \subset \mathcal{H}_1^{\mathrm{comp}}(\beta)$ there are no such saddle connections (closed geodesics), so $\hat{\chi}_\varepsilon(S) = 0$. For surfaces $S$ from the subset $\mathcal{H}_1^{\varepsilon,\mathrm{thick}}(\mathcal{C})$ there is exactly one collection like this, $\hat{\chi}_\varepsilon(S) = 1$. Finally, for the surfaces from the remaining (very small) subset $\mathcal{H}_1^{\varepsilon,\mathrm{thin}}(\mathcal{C}) = \mathcal{H}_1^{\varepsilon}(\mathcal{C}) \setminus \mathcal{H}_1^{\varepsilon,\mathrm{thick}}(\mathcal{C})$ one has $\hat{\chi}_\varepsilon(S) \geq 1$. Eskin and Masur have proved in [EM] that though $\hat{\chi}_\varepsilon(S)$ might be large on $\mathcal{H}_1^{\varepsilon,\mathrm{thin}}$ the measure of this subset is so small (see [MS]) that

$$\int_{\mathcal{H}_1^{\varepsilon,\mathrm{thin}}(\mathcal{C})} \hat{\chi}_\varepsilon(S)\, dv_1 = o(\varepsilon^2)$$

and hence

$$\int_{\mathcal{H}_1^{\mathrm{comp}}(\beta)} \hat{\chi}_\varepsilon(S)\, d\nu_1 = \mathrm{Vol}\, \mathcal{H}_1^{\varepsilon,\mathrm{thick}}(\mathcal{C}) \;+\; o(\varepsilon^2).$$

This latter volume is almost the same as the volume $\mathrm{Vol}\, \mathcal{H}_1^{\varepsilon}(\mathcal{C})$ of the neighborhood of the cusp $\mathcal{C}$ evaluated in equation (1) above, namely, $\mathrm{Vol}\, \mathcal{H}_1^{\varepsilon,\mathrm{thick}}(\mathcal{C}) = \mathrm{Vol}\, \mathcal{H}_1^{\varepsilon}(\mathcal{C}) + o(\varepsilon^2)$ (see [MS]). Taking into consideration that

$$\int_{\mathbb{R}^2} \chi_\varepsilon(x, y)\, dx\, dy = \pi\varepsilon^2$$

and applying the Siegel–Veech formula (2) to $\chi_\varepsilon$ we finally get

$$\frac{\mathrm{Vol}\, \mathcal{H}_1^{\varepsilon}(\mathcal{C})}{\mathrm{Vol}\, \mathcal{H}_1^{\mathrm{comp}}(d_1, \ldots, d_m)} + o(\varepsilon^2) = c(\mathcal{C}) \cdot \pi\varepsilon^2$$

which implies the following formula for the Siegel–Veech constant $c(\mathcal{C})$:

$$\begin{aligned} c(\mathcal{C}) &= \lim_{\varepsilon\to 0} \frac{1}{\pi\varepsilon^2} \frac{\mathrm{Vol}(\text{“}\varepsilon\text{-neighborhood of the cusp } \mathcal{C}\text{ ”})}{\mathrm{Vol}\, \mathcal{H}_1^{\mathrm{comp}}(\beta)} \\ &= (\text{explicit combinatorial factor}) \cdot \frac{\prod_{j=1}^{k} \mathrm{Vol}\, \mathcal{H}_1(\beta_k')}{\mathrm{Vol}\, \mathcal{H}_1^{\mathrm{comp}}(\beta)}. \end{aligned}$$

Sums of the Lyapunov exponents $\nu_1 + \cdots + \nu_g$ discussed in Section 1 are closely related to the Siegel–Veech constants.

## 3. Ergodic components of the Teichmüller flow

According to the theorems of H. Masur [M1] and of W. Veech [V1] Teichmüller geodesic flow is ergodic on every connected component of every stratum of flat surfaces. Thus, the Lyapunov exponents $1 + \nu_j$ of the Teichmüller geodesic flow responsible for the deviation spectrum of generic geodesics on a flat surface (see Section 1), or Siegel–Veech constants responsible for counting of closed geodesics on a flat surface (see Section 2) are specific for each connected component of each stratum. The fact that the strata $\mathcal{H}_1(d_1, \ldots, d_m)$ are not necessarily connected was observed by W. Veech.

In order to formulate the classification theorem for connected components of the strata $\mathcal{H}(d_1, \ldots, d_m)$ we need to describe the classifying invariants. There are two of them: spin structure and hyperellipticity. Both notions are applicable only to part of the strata: flat surfaces from the strata $\mathcal{H}(2d_1, \ldots, 2d_m)$ have even or odd spin structure. The strata $\mathcal{H}(2g-2)$ and $\mathcal{H}(g-1, g-1)$ have a special hyperelliptic connected component.

The results of this section are based on the joint work with M. Kontsevich [KZ].

**Spin structure.** Consider a flat surface $S$ from a stratum $\mathcal{H}(2d_1, \dots, 2d_m)$. Let $\rho\colon S^1 \to S$ be a smooth closed path on $S$; here $S^1$ is a standard circle. Note that at any point of the surfaces $S$ we know where is the "direction to the North". Hence, at any point $\rho(t) = x \in S$ we can apply a compass and measure the direction of the tangent vector $\dot{x}$. Moving along our path $\rho(t)$ we make the tangent vector turn in the compass. Thus we get a map $G(\rho)\colon S^1 \to S^1$ from the parameter circle to the circumference of the compass. This map is called the *Gauss map*. We define the *index* $\operatorname{ind}(\rho)$ of the path $\rho$ as a degree of the corresponding Gauss map (or, in other words as the algebraic number of turns of the tangent vector around the compass) taken modulo 2.

$$\operatorname{ind}(\rho) = \deg G(\rho) \mod 2.$$

It is easy to see that $\operatorname{ind}(\rho)$ does not depend on parameterization. Moreover, it does not change under small deformations of the path. Deforming the path more drastically we may change its position with respect to conical singularities of the flat metric. Say, the initial path might go on the left of $P_k$ and its deformation might pass on the right of $P_k$. This deformation changes the $\deg G(\rho)$. However, if the cone angle at $P_k$ is of the type $2\pi(2d_k + 1)$, then $\deg G(\rho) \mod 2$ does not change! This observation explains why $\operatorname{ind}(\rho)$ is well-defined for a free homotopy class $[\rho]$ when $S \in \mathcal{H}(2d_1, \dots, 2d_m)$ (and hence, when all cone angles are odd multiples of $2\pi$).

Consider a collection of closed smooth paths $a_1, b_1, \dots, a_g, b_g$ representing a symplectic basis of homology $H_1(S, \mathbb{Z}/2\mathbb{Z})$. We define the *parity of the spin-structure* of a flat surface $S \in \mathcal{H}(2d_1, \dots, 2d_m)$ as

$$\phi(S) = \sum_{i=1}^{g} (\operatorname{ind}(a_i) + 1)\,(\operatorname{ind}(b_i) + 1) \mod 2.$$

**Lemma.** *The value $\phi(S)$ does not depend on symplectic basis of cycles $\{a_i, b_i\}$. It does not change under continuous deformations of $S$ in $\mathcal{H}(2d_1, \dots, 2d_m)$.*

The lemma above shows that the parity of the spin structure is an invariant of connected components of strata of those Abelian differentials (equivalently, flat surfaces) which have zeroes of even degrees (equivalently, conical points with cone angles which are odd multiples of $2\pi$).

**Hyperellipticity.** A flat surface $S$ may have a symmetry; one specific family of such flat surfaces, which are "more symmetric than others" is of a special interest for us. Recall that there is a one-to-one correspondence between flat surfaces and pairs (Riemann surface $M$ , holomorphic 1-form $\omega$). When the corresponding Riemann surface is hyperelliptic the hyperelliptic involution $\tau\colon M \to M$ acts on any holomorphic 1-form $\omega$ as $\tau^*\omega = -\omega$.

We say that a flat surface $S$ is a hyperelliptic flat surface if there is an isometry $\tau\colon S \to S$ such that $\tau$ is an involution, $\tau \circ \tau = \mathrm{id}$, and the quotient surface $S/\tau$

is a topological sphere. In flat coordinates differential of such involution obviously satisfies $D\tau = -\mathrm{Id}$.

In a general stratum $\mathcal{H}(d_1, \ldots, d_m)$ hyperelliptic flat surfaces form a small subspace of nontrivial codimension. However, there are two special strata, namely, $\mathcal{H}(2g-2)$ and $\mathcal{H}(g-1, g-1)$, for which hyperelliptic surfaces form entire hyperelliptic connected components $\mathcal{H}^{\mathrm{hyp}}(2g-2)$ and $\mathcal{H}^{\mathrm{hyp}}(g-1, g-1)$ correspondingly.

**Remark.** Note that in the stratum $\mathcal{H}(g-1, g-1)$ there are hyperelliptic flat surfaces of two different types. A hyperelliptic involution $\tau S \to S$ may fix the conical points or might interchange them. It is not difficult to show that for flat surfaces from the connected component $\mathcal{H}^{\mathrm{hyp}}(g-1, g-1)$ the hyperelliptic involution interchanges the conical singularities.

The remaining family of those hyperelliptic flat surfaces in $\mathcal{H}(g-1, g-1)$, for which the hyperelliptic involution keeps the saddle points fixed, forms a subspace of nontrivial codimension in the complement $\mathcal{H}(g-1, g-1) \setminus \mathcal{H}^{\mathrm{hyp}}(g-1, g-1)$. Thus, the hyperelliptic connected component $\mathcal{H}^{\mathrm{hyp}}(g-1, g-1)$ does not coincide with the space of all hyperelliptic flat surfaces.

**Classification theorem for Abelian differentials.** Now, having introduced the classifying invariants we can present the classification of connected components of strata of flat surfaces (equivalently, of strata of Abelian differentials).

**Theorem 3** (M. Kontsevich and A. Zorich). *All connected components of any stratum of flat surfaces of genus $g \geq 4$ are described by the following list:*

- *The stratum $\mathcal{H}(2g-2)$ has three connected components: the hyperelliptic one, $\mathcal{H}^{\mathrm{hyp}}(2g-2)$, and two nonhyperelliptic components: $\mathcal{H}^{\mathrm{even}}(2g-2)$ and $\mathcal{H}^{\mathrm{odd}}(2g-2)$ corresponding to even and odd spin structures.*

- *The stratum $\mathcal{H}(2d, 2d)$, $d \geq 2$ has three connected components: the hyperelliptic one, $\mathcal{H}^{\mathrm{hyp}}(2d, 2d)$, and two nonhyperelliptic components: $\mathcal{H}^{\mathrm{even}}(2d, 2d)$ and $\mathcal{H}^{\mathrm{odd}}(2d, 2d)$.*

- *All the other strata of the form $\mathcal{H}(2d_1, \ldots, 2d_m)$ have two connected components: $\mathcal{H}^{\mathrm{even}}(2d_1, \ldots, 2d_m)$ and $\mathcal{H}^{\mathrm{odd}}(2d_1, \ldots, 2d_n)$, corresponding to even and odd spin structures.*

- *The stratum $\mathcal{H}(2d-1, 2d-1)$, $d \geq 2$, has two connected components; one of them: $\mathcal{H}^{\mathrm{hyp}}(2d-1, 2d-1)$ is hyperelliptic; the other $\mathcal{H}^{\mathrm{nonhyp}}(2d-1, 2d-1)$ is not.*

*All the other strata of flat surfaces of genera $g \geq 4$ are nonempty and connected.*

In the case of small genera $1 \leq g \leq 3$ some components are missing in comparison with the general case.

**Theorem 3′.** *The moduli space of flat surfaces of genus $g = 2$ contains two strata: $\mathcal{H}(1,1)$ and $\mathcal{H}(2)$. Each of them is connected and coincides with its hyperelliptic component.*

*Each of the strata $\mathcal{H}(2,2)$, $\mathcal{H}(4)$ of the moduli space of flat surfaces of genus $g = 3$ has two connected components: the hyperelliptic one, and one having odd spin structure. The other strata are connected for genus $g = 3$.*

Since there is a one-to-one correspondence between connected components of the strata and extended Rauzy classes, the classification theorem above classifies also the extended Rauzy classes.

Connected components of the strata $\mathcal{Q}(d_1, \dots, d_m)$ of meromorphic quadratic differentials with at most simple poles are classified in the paper of E. Lanneau [L].

**Bibliographical notes.** As a much more serious accessible introduction to Teichmüller dynamics I can recommend a collection of surveys of A. Eskin [E], G. Forni [Fo2], P. Hubert and T. Schmidt [HSc] and H. Masur [M2], organized as a chapter of the Handbook of Dynamical Systems. I also recommend recent surveys of H. Masur and S. Tabachnikov [MT] and of J. Smillie [S] especially in the aspects related to billiards in polygons. The part concerning renormalization and interval exchange transformations is presented in the survey of J.-C. Yoccoz [Y]. The ideas presented in the current paper are illustrated in more detailed way in the survey [Z4].

**Acknowledgements.** A considerable part of the results presented in this survey is obtained in collaboration. I use this opportunity to thank A. Eskin, M. Kontsevich and H. Masur for the pleasure to work with them. I am grateful to M.-C. Vergne for her help with the preparation of the pictures.

## References

[AEZ] Athreya, J., Eskin, A., Zorich, A., Rectangular billiards and volumes of spaces of quadratic differentials. In preparation.

[AvVi] Avila, A., Viana, M., Simplicity of Lyapunov spectra: proof of the Zorich–Kontsevich conjecture. Eprint math.DS/0508508, 2005, 36 pp.

[BMö] Bouw, I., and Möller, M., Teichmüller curves, triangle groups, and Lyapunov exponents. Eprint math.AG/0511738, 2005, 30 pp.

[Ca] Calta, K., Veech surfaces and complete periodicity in genus two. *J. Amer. Math. Soc.* **17** (2004), 871–908.

[E] Eskin, A., Counting problems in moduli space. In *Handbook of Dynamical Systems, Vol. 1B* (ed. by B. Hasselblatt and A. Katok), Elsevier Science B.V., Amsterdam 2006, 581–595.

[EM] Eskin, A., Masur, H., Asymptotic formulas on flat surfaces. *Ergodic Theory Dynam. Systems* **21** (2) (2001), 443–478.

[EMZ] Eskin, A., Masur, H., Zorich, A., Moduli spaces of Abelian differentials: the principal boundary, counting problems, and the Siegel–Veech constants. *Publ. Math. Inst. Hautes Études Sci.* **97** (1) (2003), 61–179.

[EO] Eskin, A., Okounkov, A., Asymptotics of number of branched coverings of a torus and volumes of moduli spaces of holomorphic differentials. *Invent. Math.* **145** (1) (2001), 59–104.

[FlFo] Flaminio, L., Forni, G., Invariant distributions and time averages for horocycle flows. *Duke Math. J.* **119** (3) (2003), 465–526.

[Fo1] Forni, G., Deviation of ergodic averages for area-preserving flows on surfaces of higher genus. *Ann. of Math.* (2) **155** (1) (2002), 1–103.

[Fo2] Forni, G., On the Lyapunov exponents of the Kontsevich–Zorich cocycle. In *Handbook of Dynamical Systems, Vol. 1B* (ed. by B. Hasselblatt and A. Katok), Elsevier Science B.V., Amsterdam 2006, 549–580.

[HSc] Hubert, P., and Schmidt, T., An introduction to Veech surfaces. In *Handbook of Dynamical Systems, Vol. 1B* (ed. by B. Hasselblatt and A. Katok), Elsevier Science B.V., Amsterdam 2006, 501–526.

[KeMS] Kerckhoff, S., Masur, H., and Smillie, J., Ergodicity of billiard flows and quadratic differentials. *Ann. of Math.* (2) **124** (1986), 293–311.

[Kr] Krikorian, P., Déviation de moyennes ergodiques, flots de Teichmüller et cocycle de Kontsevich–Zorich (d'après Forni, Kontsevich, Zorich). Séminaire Bourbaki 927, 2003; *Astérisque* **299** (2005), 59–93.

[K] Kontsevich, M., Lyapunov exponents and Hodge theory. In *The mathematical beauty of physics* (Saclay, 1996, in Honor of C. Itzykson), Adv. Ser. Math. Phys. 24, World Sci. Publishing, River Edge, NJ, 1997, 318–332.

[KZ] Kontsevich, M., Zorich, A., Connected components of the moduli spaces of Abelian differentials. *Invent. Math.* **153** (3) (2003), 631–678.

[L] Lanneau, E., Connected components of the moduli spaces of quadratic differentials. Eprint math.GT/0506136, 2005, 41pp.

[M1] Masur, H., Interval exchange transformations and measured foliations. *Ann. of Math.* (2) **115** (1982), 169–200.

[M2] Masur, H., Ergodic Theory of Translation Surfaces. In *Handbook of Dynamical Systems, Vol. 1B* (ed. by B. Hasselblatt and A. Katok), Elsevier Science B.V., Amsterdam 2006, 527–547.

[MS] Masur, H., Smillie, J., Hausdorff dimension of sets of nonergodic foliations. *Ann. of Math.* (2) **134** (1991), 455–543.

[MT] Masur, H., and Tabachnikov, S., Rational Billiards and Flat Structures. In *Handbook of Dynamical Systems, Vol. 1A* (ed. by B. Hasselblatt and A. Katok), Elsevier Science B.V., Amsterdam 2002, 1015–1089.

[MZ] Masur, H., and Zorich, A., Multiple Saddle Connections on Flat Surfaces and Principal Boundary of the Moduli Spaces of Quadratic Differentials. Eprint math.GT/0402197, 2004, 73pp.

[Mc1] McMullen, C., Billiards and Teichmüller curves on Hilbert modular surfaces. *J. Amer. Math. Soc.* **16** (4) (2003), 857–885.

[Mc2] McMullen, C., Dynamics of $SL_2(\mathbb{R})$ over moduli space in genus two. *Ann. of Math.*, to appear.

[Ra] Rauzy, G., Echanges d'intervalles et transformations induites. *Acta Arith.* **34** (1979), 315–328.

[S] Smillie, J., The dynamics of billiard flows in rational polygons. In *Dynamical systems, ergodic theory and applications* (ed. by Ya. G. Sinai), Encyclopaedia Math. Sci. 100, Math. Physics 1, Springer-Verlag, Berlin 2000, 360–382.

[V1] Veech, W. A., Gauss measures for transformations on the space of interval exchange maps. *Ann. of Math.* (2) **115** (1982), 201–242.

[V2] Veech, W. A., The Teichmüller geodesic flow. *Ann. of Math.* (2) **124** (1986), 441–530.

[V3] Veech, W. A., Flat surfaces. *Amer. J. Math.* **115** (1993), 589–689.

[V4] Veech, W. A., Siegel measures. *Ann. of Math.* (2) **148** (1998), 895–944.

[Vo] Vorobets, Ya., Periodic geodesics on generic translation surfaces. In *Algebraic and Topological Dynamics* (ed. by S. Kolyada, Yu. I. Manin and T. Ward), Contemp. Math. 385, Amer. Math. Soc., Providence, RI, 2005, 205–258.

[Y] Yoccoz, J.-C., Continuous fraction algorithms for interval exchange maps: an introduction. In *Frontiers in Number Theory, Physics and Geometry* (Proceedings of Les Houches winter school 2003), Volume 1, Springer-Verlag, Berlin 2006, 403–437.

[Z1] Zorich, A., Finite Gauss measure on the space of interval exchange transformations. Lyapunov exponents. *Ann. Inst. Fourier* (*Grenoble*) **46** (2) (1996), 325–370.

[Z2] Zorich, A., Deviation for interval exchange transformations. *Ergodic Theory Dynam. Systems* **17** (1997), 1477–1499.

[Z3] Zorich, A., How do the leaves of a closed 1-form wind around a surface? In *Pseudoperiodic Topology* (ed. by V. I. Arnold, M. Kontsevich, A. Zorich), Amer. Math. Soc. Trans. Ser. 2 197, Amer. Math. Soc., Providence, RI, 1999, 135–178.

[Z4] Zorich, A., Flat surfaces. In *Frontiers in Number Theory, Physics and Geometry* (Proceedings of Les Houches winter school 2003), Volume 1, Springer-Verlag, Berlin 2006, 439–585.

IRMAR, Université Rennes 1, Campus de Beaulieu, 35042 Rennes, France
E-mail: Anton.Zorich@univ-rennes1.fr

# Asymptotic behavior of smooth solutions for partially dissipative hyperbolic systems and relaxation approximation

Stefano Bianchini

**Abstract.** We study two problems related to hyperbolic systems with a dissipative source.

In the first part, we consider the asymptotic time behavior of global smooth solutions to general entropy dissipative hyperbolic systems of balance law in $m$ space dimensions, under a coupling condition among hyperbolic and dissipative part known as the Shizuta–Kawashima condition. Under the assumption of small initial data, these solutions approach constant equilibrium state in the $L^p$-norm at a rate $O\big(t^{-\frac{m}{2}(1-\frac{1}{p})}\big)$, as $t \to \infty$, for $p \in [\min\{m, 2\}, \infty]$. The main tool is given by a detailed analysis of the Green function for the linearized problem. If the space dimension $m = 1$ or the system is rotational invariant, it is possible to give an explicit form to the main terms in the Green kernel.

In the second part, we consider the hyperbolic limit of special systems of balance laws: this means to study the limit of the solution to a system of balance laws under the rescaling $(t, x) \mapsto (t/\varepsilon, x/\varepsilon)$, as $\varepsilon \to 0$. For some special dissipative systems in one space dimension, it is possible to prove the existence of the limit and to identify it as a solution to a system of conservation laws.

**Mathematics Subject Classification (2000).** 35L65.

**Keywords.** Dissipative hyperbolic systems, large time behavior, convex entropy functions, relaxation systems, Shizuta–Kawashima condition, BGK models.

## 1. Asymptotic behavior of smooth solutions to balance laws

We consider the Cauchy problem for a general hyperbolic symmetrizable $m$-dimensional system of balance laws

$$u_t + \sum_{\alpha=1}^{m} (f_\alpha(u))_{x_\alpha} = g(u), \tag{1.1}$$

with the initial conditions

$$u(x, 0) = u_0(x), \tag{1.2}$$

where $u = (u_1, u_2) \in \Omega \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with $n_1 + n_2 = n$. We also assume that there are $n_1$ conservation laws in the system, namely that we can take

$$g(u) = \begin{pmatrix} 0 \\ q(u) \end{pmatrix} \quad \text{with } q(u) \in \mathbb{R}^{n_2}. \tag{1.3}$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

According to the general theory of hyperbolic systems of balance laws [11], if the flux functions $f_\alpha$ and the source term $g$ are smooth enough, it is well known that problem (1.1)–(1.2) has a unique local smooth solution, at least for some time interval $[0, T)$ with $T > 0$, if the initial data are also sufficiently smooth. In the general case, and even for very good initial data, smooth solutions may break down in finite time, due to the appearance of singularities, either discontinuities or blow-up in $L^\infty$.

The easiest example is Burgers equation with an initial data strictly decreasing in some interval. With more generality, if we have a system of balance laws, it may happen that the source terms have no influence on the mechanism generating a singularity, or it may also help the solution to blow up.

Despite these general considerations, sometimes dissipative mechanisms due to the source term can prevent the formation of singularities, at least for some restricted classes of initial data, as observed for many models which arise to describe physical phenomena. A typical and well-known example is given by the compressible Euler equations with damping, see [20], [14] for the 1-dimensional case and [23] for an interesting 3-dimensional extension.

Recently, in [13], it was proposed a quite general framework of sufficient conditions which guarantee the global existence in time of smooth solutions. Actually, for the systems which are endowed with a strictly convex entropy function $\mathcal{E} = \mathcal{E}(u)$, a first natural assumption is the *entropy dissipation condition*, see [10], [19], [21], [24], namely for every $u, \bar{u} \in \Omega$, with $g(\bar{u}) = 0$,

$$\big(\nabla \mathcal{E}(u) - \nabla \mathcal{E}(\bar{u})\big) \cdot g(u) \leq 0,$$

where $\mathcal{E}'(u)$ is considered as a vector in $\mathbb{R}^n$ and "$\cdot$" is the scalar product in the same space.

Roughly speaking, the above condition means that the source is dissipative in some integral norm, typically $L^2$. Thus one expects that blow up in $L^\infty$ could be prevented, or certainly it does not happen for space independent solutions.

Unfortunately, it is easy to see that this condition is too weak to prevent the formation of singularities: just consider the system

$$\begin{cases} u_t + uu_x = 0, \\ v_t = -v \end{cases} \tag{1.4}$$

with entropy $u^2 + v^2$. The key point in this system is that the dissipative source $(0, -v)$ is not acting on the first equation, so that it cannot prevent the shock formation.

A quite natural supplementary condition can be imposed to entropy dissipative systems, following the classical approach by Shizuta and Kawashima [16], [22], and in the following called condition (SK), which in the present case reads

$$\mathrm{Ker}(Dg(\bar{u})) \cap \Big\{\text{eigenspaces of } \textstyle\sum_{\alpha=1}^m Df_\alpha(\bar{u})\xi_\alpha\Big\} = \{0\}, \tag{1.5}$$

for every $\xi \in \mathbb{R}^m \setminus \{0\}$ and every $\bar{u} \in \Omega$, with $g(\bar{u}) = 0$. It is possible to prove that this condition, which is satisfied in many interesting examples, is also sufficient to

establish a general result of global existence for small perturbations of equilibrium constant states.

As an example, it is easy to see that (1.4) does not satisfy (SK) condition, while the system

$$\begin{cases} u_t + uu_x + v_x = 0, \\ v_t + u_x = -v \end{cases} \tag{1.6}$$

fulfills requirement (1.5).

We investigate the asymptotic behavior in time of the global solutions, always assuming the existence of a strictly convex entropy and the (SK) condition.

Our starting point is a careful and refined analysis of the behavior of the Green function for the linearized problem around an equilibrium state $\bar{u}$,

$$u_t + \sum_{\alpha=1}^{m} A_\alpha \partial_{x_\alpha} u = Bu, \quad A^\alpha = \nabla f^\alpha(u)|_{u=\bar{u}},\ B = \nabla g(u)|_{u=\bar{u}} \tag{1.7}$$

The conditions on the existence of a dissipative strictly convex entropy and (SK) condition (and also (1.3) implies that

1. the matrices $A_\alpha$ are symmetric and

$$B = \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix}, \quad D \in \mathbb{R}^{n_2 \times n_2}, \tag{1.8}$$

   with $D$ strictly negative definite;

2. no eigenvectors of $\sum_\alpha \xi_\alpha A_\alpha$ are in the null space of $B$ for all $\xi \in \mathbb{R}^m$.

It is possible to show that the Green kernel $\Gamma(t)$ can be written as the sum of the kernels

$$\Gamma(t) = K(t) + \mathcal{K}(t). \tag{1.9}$$

The first term corresponds to a uniformly parabolic (pseudo) differential operator, while the first satisfies a uniform exponential decay in $L^2$:

$$\|D^\beta \mathcal{K}(t) w^0\|_{L^2} \le Ce^{-ct} \|D^\beta w^0\|_{L^2}. \tag{1.10}$$

It can be also shown that $\Gamma(t)$ has bounded support, so that both $K(t)$, $\mathcal{K}(t)$ have bounded support.

For the term $K(t)$ it is possible to give a more precise description, by using the two projectors $Q_0$, $Q_- = I - Q_0$: $Q_0$ is the projector on the conservative part of (1.7), i.e.

$$Q_0 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \quad Q_- \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ u_2 \end{pmatrix}.$$

Writing then

$$K(t) = \begin{bmatrix} K_{11}(t) & K_{12}(t) \\ K_{21}(t) & K_{22}(t) \end{bmatrix}, \tag{1.11}$$

it is possible to prove that $K_{11}$ decays as a heat kernel, while the other three decays as a derivative of the heat kernel:

$$\begin{aligned}
\|D^\beta K_{11}(t)w\|_{L^p} &\le C(|\beta|)\min\left\{1, t^{-\frac{m}{2}(1-\frac{1}{p})-|\beta|/2}\right\}\|w\|_{L^1},\\
\|D^\beta K_{12}(t)w\|_{L^p},\, \|D^\beta K_{21}(t)w\|_{L^p} &\le C(|\beta|)\min\left\{1, t^{-\frac{m}{2}(1-\frac{1}{p})-1/2-|\beta|/2}\right\}\|w\|_{L^1},\\
\|D^\beta K_{22}(t)w\|_{L^p} &\le C(|\beta|)\min\left\{1, t^{-\frac{m}{2}(1-\frac{1}{p})-1-|\beta|/2}\right\}\|w\|_{L^1}.
\end{aligned} \tag{1.12}$$

Using this Green kernel representation, one can show that solution to (1.1) satisfies the same decay estimates as $\Gamma(t)$. More precisely, one can prove that

$$\|u(t) - K(t)Q_0u(0)\|_{L^p} \le C\min\left\{1, t^{-\frac{m}{2}(1-\frac{1}{p})-1/2}\right\}\|u(0)\|_{L^1}.$$

Clearly to have more information on the asymptotic behavior of $u(t)$ one needs to know more information on $K(t)$. This can be done in two situations: if the space dimension is $m = 1$, or under the assumption of rotational invariance. In both cases the Fourier components of the differential operators can be inverted.

As an example, we can consider the linearized isentropic Euler equations with damping,

$$\begin{cases} \rho_t + \operatorname{div} v = 0, \\ v_t + \nabla\rho = -v. \end{cases} \tag{1.13}$$

One can check that the three conditions are satisfied. We can decompose the Green kernel $\Gamma$ in three parts

$$\Gamma(t,x) = K(t,x) + R(t,x) + \mathcal{K}(t,x), \tag{1.14}$$

where $K(t,x)$ can be computed to be

$$K(t,x) = \begin{bmatrix} G(t,x) & (\nabla G(t,x))^T \\ \nabla G(t,x) & \nabla^2 G(t,x) \end{bmatrix} + R_1(t,x), \tag{1.15}$$

where $G(t,x)$ is the heat kernel for $u_t = \Delta u$, and the rest term $R_1(t,x)$ satisfies the bound

$$R_1(t,x) = \frac{e^{-c|x|^2/t}}{(1+t)^2}\begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(1)(1+t)^{-1/2} \\ \mathcal{O}(1)(1+t)^{-1/2} & \mathcal{O}(1)(1+t)^{-1} \end{bmatrix}. \tag{1.16}$$

In particular the principal part of $\Gamma(t)$ is given by the heat kernel $G(t,x)$.

The rest part $R_1$ is exponentially decreasing and smooth, while $\mathcal{K}(t,x)$ can be computed to be

$$\mathcal{K}(t,x) = \begin{bmatrix} 0 & 0 \\ 0 & e^{-t}\mathcal{P} \end{bmatrix} + e^{-t}\begin{bmatrix} W_{00}(t,x) & W_{01}(t,x) \\ W_{10}(t,x) & W_{11}(t,x) \end{bmatrix} + R_2(t,x). \tag{1.17}$$

Here $\mathcal{P}: (L^2(\mathbb{R}^3))^3 \mapsto (L^2(\mathbb{R}^3))^3$ is the orthogonal projection of $L^2$ vector fields on the subspace of divergence free vector fields. $\mathcal{P}v$ is characterized by

$$\mathcal{P}v \in (L^2(\mathbb{R}^3))^3, \quad \mathrm{div}\mathcal{P}v = 0, \quad \mathrm{curl}(v - \mathcal{P}v) = 0,$$

and so we have that

$$v - \mathcal{P}v = \nabla\psi \quad \text{with } \Delta\psi = \mathrm{div}v.$$

This yields

$$\mathcal{P}v = v - \nabla(\Delta^{-1}\mathrm{div}v). \tag{1.18}$$

In fact, in Fourier coordinates, we have

$$\widehat{\mathcal{P}v}(\xi) = \hat{v}(\xi) - |\xi|^{-2}(\xi \cdot \hat{v}(\xi))\xi = \hat{v}(\xi) - |\xi|^{-2}\xi\xi^T \cdot \hat{v}(\xi). \tag{1.19}$$

The matrix valued function

$$W(t,x) = W_1(t,x) + W_2(t,x) = \begin{bmatrix} W_{00}(t,x) & W_{01}(t,x) \\ W_{10}(t,x) & W_{11}(t,x) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \delta(x)\mathcal{P} \end{bmatrix} \tag{1.20}$$

is the matrix valued Green function of the system

$$\begin{cases} \rho_t + \mathrm{div}v = 0, \\ v_t + \nabla\rho = 0, \end{cases}$$

and it can be written by means of the fundamental solution to the wave equation $u_{tt} = \Delta u$. In fact, $W_{00}$ is the solution of $u_{tt} = \Delta u$ with initial data $u = \delta(x)$, $u_t = 0$, and

$$W_1 = \begin{bmatrix} W_{00} & \nabla^T\partial_t(-\Delta)^{-1}W_{00} \\ \nabla\partial_t(-\Delta)^{-1}W_{00} & -\nabla^2(-\Delta)^{-1}W_{00} \end{bmatrix}. \tag{1.21}$$

In particular one can check that $W_2$ corresponds to incompressible vector fields, while $W_1$ corresponds to curl free vector fields.

An interesting open question is whether the (SK) can be relaxed by considering it acting only on a subset of the eigenspaces of $\sum Df_\alpha\xi_\alpha$. As a trivial example, if $f^\alpha$ are linear functions, then no dissipativity condition is needed. In [27] it is possible to find a non trivial example of a system which does not satisfy (SK) condition but still the smooth solution exists for all $t \geq 0$.

## 2. Relaxation limit of balance laws

After having proved existence for small data of the system

$$u_t + \sum_{\alpha=1}^{m} (f_\alpha(u))_{x_\alpha} = \frac{g(u)}{\varepsilon}, \tag{2.1}$$

under the dissipativity and (SK) condition, one can ask the question of describing the hyperbolic limit as $\varepsilon \to 0$. At a formal level, by writing

$$g(u) = g(u_1, u_2) = 0 \quad \Longrightarrow \quad u_1 = q(u_2) \tag{2.2}$$

for the equilibrium manifold, at a formal level we obtain the symmetrizable hyperbolic system

$$u_{1,t} + \sum_{\alpha=1}^{m} \left(f_\alpha(u_1, q(u_1))\right)_{x_\alpha} = 0. \tag{2.3}$$

As an example, if (2.1) is a kinetic scheme, then this limit corresponds to the hydrodynamic limit, i.e. to the rescaling $(t, x) \mapsto (t/\varepsilon, x/\varepsilon)$. The most outstanding problem in this direction is the hydrodynamic limit of Boltzmann equation, at least in one space dimension. In this direction the major result is the convergence under the assumption of piecewise smooth solution to the limiting Euler system with non interacting shocks [26]. We remark that hyperbolic limit must be based on the proof of local existence for data which do not satisfy any integral estimates.

We want to underline the difference among the known results and the theory of hyperbolic system. The assumption of non interacting shocks is equivalent to saying that we do not need to control the non linear interaction among waves, which is one of the key aspects of hyperbolic systems.

The main tool for proving existence and stability of solutions to hyperbolic systems is the local decomposition of the solution into waves and the description of their interaction. As an example, for BV solutions of an $n \times n$ system, the derivative of the solution is decomposed in $n$ scalar measures, each measure representing the waves of one of the characteristic families of $u$. The key point is that one can describe the evolution of these waves as they interact and find an interaction functional which bounds the total interaction.

We believe that this functional would be a key point in proving hyperbolic limit of balance laws. It is clear that understanding this functional implies the understanding of the wave structure of the solution.

In general it seems difficult to prove convergence of (2.1) as $\varepsilon \to 0$, under the assumptions considered in the previous part. We thus restrict to particular quasilinear systems of the form (BGK schemes with one moment) in one space dimension:

$$\partial_t F^\alpha + \alpha F^\alpha_x = \frac{1}{\varepsilon}\big(M^\alpha(u) - F^\alpha\big), \quad u = \sum_\alpha F^\alpha. \tag{2.4}$$

The functions $M^\alpha(u)$ are the Maxwellians. At a formal level, as $\varepsilon \to 0$ one obtains

$$F^\alpha = M^\alpha(u), \quad u_t + \mathcal{F}(u)_x = 0,$$

where the flux function $\mathcal{F}(u)$ is given by

$$\mathcal{F}(u) = \sum_\alpha \alpha M^\alpha(u).$$

The easiest example is the scheme introduced in [15],

$$\begin{cases} u_t + v_x = 0, \\ v_t + \Lambda^2 u_x = \dfrac{1}{\varepsilon}(\mathcal{F}(u) - v), \end{cases} \tag{2.5}$$

which can be put in the form (2.4) by diagonalizing the semilinear part.

By differentiating the second equation of (2.5) w.r.t. $x$ and using the first one obtains the nonlinear wave equation

$$u_t + A(u)u_x = \frac{1}{\varepsilon}(u_{xx} - u_{tt}), \tag{2.6}$$

with $A(u) = D\mathcal{F}(u)$. The above equation is meaningful also in the case $A(u)$ is not a Jacobian matrix, so that one cannot write a conservative form like (2.5). For this particular system it is now proved the existence and stability of solutions with initial data of small BV norm [3].

In this last part we describe the structure of the waves of (2.6). Writing the system in the form

$$\begin{cases} F^-_t - F^-_x = (M^-(u) - F^-)/\varepsilon, \\ F^*_t + F^*_x = (M^+(u) - F^+)/\varepsilon, \end{cases} \qquad u = F^- + F^+,\ M^{\pm}(u) = \frac{u \pm \mathcal{F}(u)}{2} \tag{2.7}$$

with $F^-$, $F^+ \in \mathbb{R}^n$, and assuming the stability condition $|D\mathcal{F}(u)| < 1$, at a formal level one expects that as $\varepsilon \to 0$, the function $u$ converges to a solution to

$$u_t + \mathcal{F}(u)_x = 0, \quad u \in \mathbb{R}^n. \tag{2.8}$$

Concerning the solution of (2.8), we know that its structure can be described as the nonlinear sum of $n$ shock waves, corresponding to the characteristic speed of $D\mathcal{F}$,

$$u_x(t,x) = \sum_{i=1}^{n} v_i(t,x)\tilde{r}_i(t,x), \quad v_i \in \mathbb{R},\ \tilde{r}_i \in \mathbb{R}^n. \tag{2.9}$$

The vectors $\tilde{r}_i$ are in general not the eigenvectors of $D\mathcal{F}$, but are close to them for small data. Also the propagation speeds of the scalar $v_i$ is close to the $i$-th eigenvalue. Their interaction is described by a Glimm type functional [4]. Thus, a possible way of thinking of the waves in (2.7) is to imagine that the solution $(F^-, F^+)$ is the sum of $n$ waves (the wave decomposition of $u$) and some remaining term $v$ which is dissipating entropy.

It turns out that a more natural description is that the solution to (2.7) can be decomposed into $2n$ waves, $n$ for each component $F^-$, $F^+$,

$$F^-_x(t,x) = \sum_{i=1}^{n} f_i^-(t,x)\tilde{r}_i^-(t,x), \quad F^+_x(t,x) = \sum_{i=1}^{n} f_i^+(t,x)\tilde{r}_i^+(t,x). \tag{2.10}$$

This in some sense is a different philosophy, because we are not describing the solution as the approximate solution to the limiting hyperbolic system plus a term which is dissipating: we give a full nonlinear structure to the solution to the kinetic scheme.

We remark that with more generality, the BGK system (2.4) is decomposed as the sum of $n$ waves for each component $F^\alpha$. It is an open question which is the right decomposition in non linear waves for the more complicated discrete models, for example the well-known Broadwell model in one space dimension.

## References


[1] Bianchini, S., BV solutions to semidiscrete schemes. *Arch. Rat. Mech. Anal.* **167** (1) (2003), 1–81.

[2] Bianchini, S., Glimm interaction functional for BGK schemes. Preprint.

[3] Bianchini, S., Hyperbolic limit of the Jin-Xin relaxation model. Preprint.

[4] Bianchini, S., and Bressan,A., Vanishing viscosity solutions of nonlinear hyperbolic systems. *Ann. of Math.* **161** (2005), 223-342.

[5] Bianchini, S., Hanouzet, B., and Natalini, R., Asymptotic Behavior of Smooth Solutions for Partially Dissipative Hyperbolic Systems with a Convex Entropy. Preprint.

[6] Bouchut, F., Construction of BGK models with a family of kinetic entropies for a given system of conservation law. *J. Statist. Phys.* **95** (1999), 113–170.

[7] Bressan, A., The unique limit of the Glimm scheme. *Arch. Rational Mech. Anal.* **130** (1995), 205–230.

[8] Bressan, A., *Hyperbolic systems of conservation laws*. Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, Oxford 2000.

[9] Bressan, A., Liu, T. P., and Yang, T., $L^1$ stability estimates for $n \times n$ conservation laws. *Arch. Rat. Mech. Anal.* **149** (1999), 1–22.

[10] Chen, G.-Q., Levermore, C. D., and Liu, T.-P., Hyperbolic conservation laws with stiff relaxation terms and entropy. *Comm. Pure Appl. Math.* **47** (1994), 787–830.

[11] Dafermos, C. M., *Hyperbolic conservation laws in continuum physics*. Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin 2000.

[12] Glimm, J., Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.* **18** (1965), 697–715.

[13] Hanouzet, B., and Natalini, R., Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy. *Arch. Ration. Mech. Anal.* **169** (2) (2003), 89–117.

[14] Hsiao, L., and Liu, T.-P., Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping. *Comm. Math. Physics* **143** (1992), 599–605.

[15] Jin, S., and Xin, Z., The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm. Pure Appl. Math.* **48** (1995), 235–276.

[16] Kawashima, S., Large-time behaviour of solutions to hyperbolic-parabolic systems of conservation laws and applications. *Proc. Roy. Soc. Edinburgh Sect. A* **106** (1–2) (1987), 169–194.

[17] Liu, T.-P., Hyperbolic conservation laws with relaxation. *Comm. Math. Phys.* **108** (1987), 153–175.

[18] Mascia, C., Zumbrun, K., Pointwise Green function bounds for shock profiles of systems with real viscosity. *Arch. Ration. Mech. Anal.* **169** (3) (2003), 177–263.

[19] Müller, I., and Ruggeri, T., *Rational extended thermodynamics*. Second ed., with supplementary chapters by H. Struchtrup and Wolf Weiss, Springer Tracts Nat. Philos. 37, Springer-Verlag, New York 1998.

[20] Nishida, T., *Nonlinear hyperbolic equations and related topics in fluid dynamics*. Département de Mathématique, Université de Paris-Sud, Orsay, 1978, Publications Mathématiques d'Orsay, No. 78-02.

[21] Ruggeri, T., The entropy principle: from continuum mechanics to hyperbolic systems of balance laws. *Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat.* (8) **8** (2005), 1–20.

[22] Shizuta, Y., and Kawashima, S., Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation. *Hokkaido Math. J.* **14** (2) (1985), 249–275.

[23] Sideris, T., Thomases, B., and Wang, D., Decay and singularities of solutions of the three-dimensional Euler equations with damping. *Comm. Partial Differential Equations* **28** (3–4) (2003), 795–816.

[24] Yong, W.-A., Basic aspects of hyperbolic relaxation systems. In *Advances in the theory of shock waves*, Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser, Boston, MA, 2001, 259–305.

[25] Yong, W.-A., Entropy and global existence for hyperbolic balance laws. *Arch. Ration. Mech. Anal.* **172** (2) (2004), 247–266.

[26] Yu, S.-H., Hydrodynamic limits with Shock Waves of the Boltzmann Equation. *Commun. Pure Appl. Math.* **58** (3) (2005), 409–443.

[27] Zeng, Y., Gas dynamics in thermal nonequilibrium and general hyperbolic systems with relaxation. *Arch. Ration. Mech. Anal.* **150** (3) (1999), 225–279.

SISSA-ISAS, Via Beirut 2–4, 34014 Trieste, Italy
E-mail: bianchin@sissa.it

# Nonlinear Schrödinger equations in inhomogeneous media: wellposedness and illposedness of the Cauchy problem

Patrick Gérard

**Abstract.** We survey recent wellposedness and illposedness results for the Cauchy problem for nonlinear Schrödinger equations in inhomogeneous media such as Riemannian manifolds or domains of the Euclidean space, trying to emphasize the influence of the geometry. The main tools are multilinear Strichartz estimates for the Schrödinger group.

**Mathematics Subject Classification (2000).** Primary 35Q55; Secondary 35B30.

**Keywords.** Nonlinear Schrödinger equations, Strichartz inequalities, multilinear estimates, eigenfunction estimates.

## 1. Introduction

The nonlinear Schrödinger equation arises in several areas of Physics (see the book [55] for an introduction), such as Optics or Quantum Mechanics, where it is related to Bose–Einstein condensation or Superfluidity. From the mathematical point of view, this equation has been studied on the Euclidean space since the seventies. However, it is quite relevant, in the above applications, to consider this equation on inhomogeneous media. In Optics, for instance, this naturally corresponds to a variable optical index; more specifically, spatial inhomogeneity has been recently used in the modelization of broad-area semiconductor lasers (see [38]). One of the main mathematical questions is then to evaluate the impact of the inhomogeneity on the dynamics of the equation, in particular regarding the wellposedness theory of the Cauchy problem. The goal of this paper is to survey recent mathematical contributions in this direction.

Let us precise what we mean by inhomogeneous medium in this context. Our physical space $M$ is either the space $\mathbb{R}^d$ or a compact manifold, endowed in both cases with a second order differential operator $P$, which is elliptic, positive and selfadjoint with respect to some Lebesgue density $\mu$ and satisfies $P(1) = 0$.[1]

In coordinates, this means that $P$ and $\mu$ are given by

$$Pu = -\frac{1}{\rho}\nabla.(A\nabla u), \quad d\mu = \rho(x)\,dx, \tag{1.1}$$

[1] Notice that this latter condition prevents potential terms in $P$. We impose this condition here for the sake of concision, though potential terms may of course be quite relevant too.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

where $\rho$ is a smooth positive function, and where $A$ is a smooth function valued in positive definite matrices. If $M = \mathbb{R}^d$, we shall impose the following additional conditions

$$0 < c \leq \rho(x), \quad cI \leq A(x), \quad |\partial^\alpha \rho(x)| + |\partial^\alpha A(x)| \leq C_\alpha, \quad \alpha \in \mathbb{N}^d, \tag{1.2}$$

in order to avoid degeneracy at infinity. An example of such an operator is of course minus the Laplace operator associated to a Riemannian metric $g$ on $M$ which, in the case $M = \mathbb{R}^d$, satisfies moreover

$$cI \leq g(x), \quad |\partial^\alpha g(x)| \leq C_\alpha, \quad \alpha \in \mathbb{N}^d.$$

In this setting, the nonlinear Schrödinger equation (NLS) reads

$$i\frac{\partial u}{\partial t} - Pu = F(u), \tag{1.3}$$

where the unknown complex function $u$ depends on $t \in \mathbb{R}$ and on $x \in M$, and the Cauchy problem consists in imposing the initial value of $u$ at $t = 0$. Here the nonlinearity $F$ is a smooth function on $\mathbb{C}$, which we assume to satisfy the following normalization and growth conditions:

$$F(0) = 0, \quad |D^k F(z)| \leq C_k(1 + |z|)^{1+\alpha-k}, \quad k = 0, 1, 2, \ldots. \tag{1.4}$$

In many situations, we require additional conditions on the structure of $F$. The most common one imposes that $F$ derives from a potential function

$$F(z) = \frac{\partial V}{\partial \bar{z}}, \quad V : \mathbb{C} \to \mathbb{R}. \tag{1.5}$$

In this case (1.3) is a Hamiltonian system with the following Hamiltonian functional:

$$H(u) = \int_M (Pu\,\bar{u} + V(u))\,d\mu, \tag{1.6}$$

and consequently it formally enjoys the conservation law

$$H(u(t)) = H(u(0)). \tag{1.7}$$

Furthermore, if we assume the following gauge-invariance condition

$$V(\mathrm{e}^{i\theta} z) = V(z), \quad \theta \in \mathbb{R}, \tag{1.8}$$

i.e. $V(z) = G(|z|^2)$, $F(z) = G'(|z|^2)z$, we also have the $L^2$ conservation law

$$\|u(t)\|_{L^2(M,\mu)} = \|u(0)\|_{L^2(M,\mu)}. \tag{1.9}$$

A typical example of nonlinearity $F$ satisfying (1.4), (1.5) and (1.8) is

$$F_{\alpha,\pm}(z) = \pm(1 + |z|^2)^{\alpha/2} z. \tag{1.10}$$

Finally, let us indicate that we shall sometimes discuss another kind of inhomogeneous NLS, namely when $-P$ is the Laplace operator with Dirichlet or Neumann boundary condition on a smooth domain of the Euclidean space. However, the theory is much less complete in this context.

This paper is organized as follows. After defining three different notions of well-posedness for the Cauchy problem for (1.3) on the scale of Sobolev spaces in Section 2, we make some general observations based on scaling considerations in Section 3. We begin Section 4 by recalling the role of Strichartz estimates in the analysis of (1.3) on the Euclidean space. We insist that this part is by no means an exhaustive review of the NLS theory on the Euclidean space. In particular, we did not discuss the recent contributions on scattering theory and on blow up. Then we really start the study of the influence of the geometry by observing that losses of derivatives may appear in Strichartz inequalities in the case of inhomogeneous media. In Section 5, we revisit the wellposedness problems by introducing multilinear Strichartz estimates, which originate in the works of Bourgain for Schrödinger and of Klainerman–Machedon for the wave equations. Finally, Section 6 is devoted to discussing in details the case of simple Riemannian compact manifolds, such as tori and spheres.

## 2. Some notions of wellposedness

We start with defining precisely the notions of wellposedness we are going to use throughout this paper. Indeed, since our evolution problem is nonlinear, several notions are available. We shall define these notions for the nonlinear Schrödinger equation (1.3) but it is clear that these notions are quite general and can be applied to other evolution equations.

**Definition 2.1.** We shall say that the Cauchy problem for equation (1.3) is (locally) *well-posed* on $H^s(M)$ if, for every bounded subset $B$ of $H^s(M)$, there exists $T > 0$ and a Banach space $X_T$ continuously contained in $C([-T, T], H^s(M))$ such that:

i) For every Cauchy data $u_0 \in B$, (1.3) has a unique solution $u \in X_T$ such that $u(0) = u_0$.

ii) If $u_0 \in H^\sigma(M)$ for $\sigma > s$, then $u \in C([-T, T], H^\sigma(M))$.

iii) The map
$$u_0 \in B \mapsto u \in X_T$$
is continuous.

Moreover, we shall say that the Cauchy problem for equation (1.3) is *globally well-posed* on $H^s(M)$ if properties i), ii), iii) above hold for every time $T > 0$.

Notice that in some cases local wellposedness can be combined with the conservation laws (1.9) and (1.7) to provide global wellposedness. Specifically, assume for instance that (1.3) is well-posed on $L^2(M) = H^0(M)$ and that $F$ is gauge-invariant.

Since the $L^2$ conservation law holds for every solution in $C([-T, T], H^s(M))$ with $s$ large enough, it results from requirements ii) and iii) that this conservation law holds on $[-T, T]$ as soon as $u_0 \in L^2$. Combining this observation with requirement i), we conclude that (local) wellposedness on $L^2$ implies global wellposedness on $L^2$. Similarly, one can show that local wellposedness on $H^1$ implies global wellposedness on $H^1$, under the assumption that a bound on $\|f\|_{L^2}$ and on $H(f)$ is equivalent to a bound on $\|f\|_{H^1}$, as it is the case, for instance, if $F$ is gauge invariant and derives from a nonnegative potential with $(d-2)\alpha \leq 4$.

**Definition 2.2.** We shall say that the Cauchy problem for equation (1.3) is (locally) *uniformly well-posed* on $H^s(M)$ if it is well-posed on $H^s(M)$ and if, with the notation of Definition 2.1 the map $u_0 \in B \mapsto u \in X_T$ is *uniformly* continuous.

One defines similarly *global uniform wellposedness*. Compared to Definition 2.1, uniform wellposedness can be understood as an additional requirement of high frequency stability for small uniform time. Let us mention that, in all the positive results of this paper, uniform continuity will come from Lipschitz continuity. As we shall see in the next sections, uniform wellposedness is rather natural for *semilinear* equations such as (1.3), but it may be violated for other natural evolution equations. For instance, it can be shown (see e.g. [59]) that the Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

is well-posed on $H^s(\mathbb{R}, \mathbb{R})$ for $s > 3/2$, but is not uniformly well-posed. This is related to the *quasilinear* hyperbolic feature of Burgers' equation. A more subtle example is the modified Korteweg–de Vries equation on the one-dimensional torus,

$$\frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} + \left(u^2 - \int_{\mathbb{T}} u^2\, dx\right)\frac{\partial u}{\partial x} = 0,$$

which is uniformly well-posed on $H^s(\mathbb{T}, \mathbb{R})$ for $s > 1/2$ (see [10]), but is not for $s \in ]3/8, 1/2[$, though it is well-posed for $s$ in this interval. (see [56] and [40]).

**Definition 2.3.** We shall say that the Cauchy problem for equation (1.3) is (locally) *regularly well-posed* on $H^s(M)$ if it is well-posed on $H^s(M)$ and if, with the notation of Definition 2.1, the map $u_0 \in \mathring{B} \mapsto u \in X_T$ is *smooth*.

One defines similarly *global regular wellposedness*. As we shall see in the next section, regular wellposedness is quite a stringent notion if $F$ is not a polynomial. On the other hand, for polynomial gauge-invariant nonlinearities, it will lead us to multilinear Strichartz estimates (see Section 5 below), which turn out to be the key estimates in this theory.

## 3. General observations

Since the free group $e^{-itP}$ acts on $H^s(M)$, the following result is an elementary consequence of the classical nonlinear estimates in Sobolev spaces $H^s$ for $s > d/2$, combined with the Duhamel formulation of the Cauchy problem for (1.3),

$$u(t) = e^{-itP}u_0 - i\int_0^t e^{-i(t-t')P}(F(u(t')))\,dt'. \tag{3.1}$$

**Proposition 3.1.** *If $s > d/2$, the Cauchy problem for* (1.3) *is regularly well-posed on $H^s(M)$.*

The above proposition has the following partial converse.

**Proposition 3.2.** *Assume $F(0) = 0$ and $D^kF(0) \neq 0$ for some $k \geq 2$ and that the Cauchy problem for* (1.3) *is regularly well-posed on $H^s(M)$. Then $s \geq \frac{d}{2} - \frac{2}{k-1}$.*

**Corollary.** *If $F(0) = 0$ and $F$ is real analytic and is not polynomial, then the Cauchy problem for* (1.3) *is not regularly well-posed on $H^s(M)$ for $s < \frac{d}{2}$.*

Proposition 3.2 relies on a very simple idea: if we solve (1.3) with the following bounded data in $H^s$,

$$u(0, x) = f_N(x) = N^{d/2-s}\varphi(Nx),$$

where $\varphi$ is a suitable cutoff function – so that the above expression makes sense on a manifold, choosing local coordinates –, and $N$ is a large parameter, then, because $P$ is a differential operator of order 2, on times $t$ such $|t| \ll N^{-2}$, the term $Pu$ can be neglected at the first order. For instance, for such times,

$$e^{-itP}f_N - f_N \to 0$$

in $H^s$ as $N$ tends to infinity. Then one uses this remark to compute successively

$$v_j(t) = \frac{d^j}{d\delta^j}(u^\delta(t))_{|\delta=0}, \quad j \geq 1,\ |t| \ll N^{-2},$$

where $u^\delta$ denotes the solution to (1.3) such that $u^\delta(0) = \delta f_N$. The continuity of the differential of order $k$ of the map $u_0 \in H^s \mapsto u(t) \in H^s$ implies

$$\|v_k(t)\|_{H^s} \leq C\|f_N\|_{H^s}^k$$

which, for $t = N^{-2-\varepsilon}$, yields the claimed condition on $s$.

Combined with more delicate estimates, this idea can also be used to disprove other kinds of wellposedness, as in the following adaptation to inhomogeneous media of a recent result by Christ, Colliander and Tao.

**Theorem 3.3** (Christ–Colliander–Tao [27], Burq–Gérard–Tzvetkov [17]). *If $\alpha > 0$, the Cauchy problem for* (1.3) *with $F = F_{\alpha,\pm}$ given by* (1.10) *is not uniformly well-posed on $H^s(M)$ if $s < 0$, and not well-posed on $H^s(M)$ if $0 < s < \frac{d}{2} - \frac{2}{\alpha}$.*

The main point of the proof is to establish that, for sufficiently small times – but not too small – , the solution of the above equation with data $\kappa_N f_N$, where $\kappa_N$ is a small coefficient to be adjusted, is approximated by the solution $v_N$ of the ordinary differential equation

$$i\partial_t v_N = \pm(1 + |v_N|^2)^{\alpha/2} v_N$$

with the same Cauchy data. Of course $v_N$ can be computed explicitly,

$$v_N(t, x) = \mathrm{e}^{\mp it(1+\kappa_N^2 |f_N(x)|^2)^{\alpha/2}} \kappa_N f_N(x),$$

and one checks that the above oscillating term induces instability in the first case, while in second case it produces norm inflation, namely the $H^s$ norm of the solution can become unbounded for a sequence of times tending to 0, though the Cauchy data tend to 0 in $H^s$.

## 4. The role of Strichartz inequalities

In this section we first recall the basic role played by Strichartz inequalities in the analysis of equation (1.3) on Euclidean spaces, quoting some important results in this context, without pretending to be exhaustive. Then we discuss extensions of these inequalities to different geometries.

**4.1. The Euclidean case.** In this subsection we assume that $M = \mathbb{R}^d$ and that $-P$ is the Laplace operator. In this case, the solution of the linear Schrödinger equation is explicit,

$$\mathrm{e}^{it\Delta} u_0(x) = \frac{1}{(4i\pi t)^{d/2}} \int_{\mathbb{R}^d} \mathrm{e}^{i|x-y|^2/4t} u_0(y)\, dy, \tag{4.1}$$

and this implies the following dispersion estimate:

$$\|\mathrm{e}^{it\Delta} u_0\|_{L^\infty(\mathbb{R}^d)} \leq \frac{1}{(4\pi|t|)^{d/2}} \|u_0\|_{L^1(\mathbb{R}^d)}. \tag{4.2}$$

By a classical functional-analytic tool (known as the $TT^*$ trick), this estimate implies important inequalities for the solution of the linear Schrödinger equation with $L^2$ data. In order to state these inequalities we shall say that a pair $(p, q) \in [1, \infty] \times [1, \infty]$ is $d$-admissible if

$$\frac{2}{p} + \frac{d}{q} = \frac{d}{2}, \quad p \geq 2, \ (p, q) \neq (2, \infty). \tag{4.3}$$

Moreover, if $r \in [1, \infty]$, we denote by $\bar{r}$ the conjugate exponent of $r$, characterized by

$$\frac{1}{\bar{r}} + \frac{1}{r} = 1.$$

**Proposition 4.1** (Strichartz [54], Ginibre–Velo [34], Yajima [61], Keel–Tao [42]). *Let $(p_1, q_1)$, $(p_2, q_2)$ be $d$-admissible pairs. There exists $C > 0$ such that, for every $u_0 \in L^2(\mathbb{R}^d)$, for every $T > 0$ and $f \in L^{\bar{p}_1}([0, T], L^{\bar{q}_1}(\mathbb{R}^d))$, the solution $u$ of*

$$i\partial_t u + \Delta u = f, \quad u(0) = u_0,$$

*satisfies $u \in L^{p_2}([0, T], L^{q_2}(\mathbb{R}^d))$ with the inequality*

$$\|u\|_{L^{p_2}([0,T],L^{q_2}(\mathbb{R}^d))} \leq C\,(\|u_0\|_{L^2(\mathbb{R}^d)} + \|f\|_{L^{\bar{p}_1}([0,T],L^{\bar{q}_1}(\mathbb{R}^d))}). \tag{4.4}$$

These inequalities were first proved in [54] in the particular cases $p_1 = q_1$ and $p_2 = q_2$, then in [34] in the cases $p_1 = p_2 > 2$, then in [61] for $p_1$, $p_2 > 2$ arbitrary, and finally, for the endpoint case $p = 2$, in [42], where an abstract presentation of the $TT^*$ trick is also available. Notice that these estimates are optimal: indeed, the first condition in (4.3) is related to the scale invariance of the Schrödinger equation, the second condition comes from general properties of translation invariant operators on $L^p$ (see *e.g.* Theorem 1.1 of Hörmander [39]), while the special forbidden case $p = 2, q = \infty$, which only arises for $d = 2$, has been checked by Montgomery-Smith [46]. A variant of the proof of the necessity of the scaling condition consists in testing the above estimates for $f = 0$ and

$$u_0(x) = \varphi(Nx),$$

where $\varphi \in C_0^\infty(\mathbb{R}^d)$ and $N$ is a large parameter. As we already observed in the previous section, for $t \ll N^{-2}$ the solution $u(t, x)$ stays essentially constant. This is made more precise by the ansatz

$$u(t, x) \sim \psi(N^2 t, Nx)$$

where $\psi(s, .)$ belongs to the Schwartz class, uniformly as $s$ stays bounded, and $\psi(0, x) = \varphi(x)$. Consequently, if $(p, q)$ is an admissible pair, then the $L_t^p(L_x^q)$ norm of $u$ on the thin slab $[0, N^{-2}] \times \mathbb{R}^d$ is equivalent to $N^{-2/p-d/q} = N^{-d/2}$, which is the magnitude of the $L^2$ norm of $u_0$. In other words, the main contribution in the $L_t^p(L_x^q)$ norm of $u$ already lies in the thin slab $[0, N^{-2}] \times \mathbb{R}^d$: this is a striking illustration of the dispersive character of the Schrödinger equation. Moreover, this remark can be carried out to inhomogeneous media, showing that *the above Strichartz inequalities cannot be improved in any inhomogeneous media.* On the other hand, as we shall see in the sequel, they may be dramatically altered.

We close this subsection by pointing that Strichartz inequalities on the Euclidean space essentially provide reverse statements of the necessary conditions to wellposedness of Proposition 3.2 and Theorem 3.3.

**Theorem 4.2** (Ginibre–Velo [33], [34], Kato [41], Cazenave–Weissler [24], Tsutsumi [58], Yajima [61]). *Let $F$ satisfy conditions* (1.4) *and let $s \geq 0$ satisfy*

$$s > \frac{d}{2} - \frac{2}{\alpha}.$$

*Then equation* (1.3) *is uniformly well-posed on $H^s(\mathbb{R}^d)$ if $d \leq 6$. Moreover, if $F$ is a polynomial of degree $1+\alpha$, then* (1.3) *is regularly well-posed on $H^s(\mathbb{R}^d)$ for every $d$.*

The unexpected condition $d \leq 6$ relies on lack of good estimates for large derivatives of $F(u)$ if $F$ is non-polynomial. This limitation may not be optimal, but so far it cannot be dropped, for instance for the proof of propagation of high regularity.

In the polynomial case, such a condition is not necessary, and we observe that, under the additional conditions (1.5) and (1.8), thresholds of the three wellposedness properties coincide, for every $d$, with $\max(0, d/2 - 2/\alpha)$. Furthermore, let us mention that some regular wellposedness results were proved on $H^s$ for $s < 0$, in the one-dimensional case, for quadratic nonlinearities which do not satisfy the gauge invariance condition (see Kenig–Ponce–Vega [43]).

The critical cases $s_c = d/2 - 2/\alpha \geq 0$ are not covered by the above theorem. Using the same Strichartz inequalities, it is possible to extend the wellposedness results of this theorem on a global time interval *for data which lie in a small neighborhood of* 0 *in $H^s$* (see [24]). For large data however, this question is still the object of intensive work, particularly for Hamiltonian gauge-invariant nonlinearities and $s_c = 0$ or 1. For instance, if $d = 2$ and $F(u) = -|u|^2 u$, $s_c = 0$ and the existence of blow up solutions (see Zakharov [62]) together with a scaling argument yields a family of solutions with bounded data in $L^2(\mathbb{R}^2)$ and which blow up at arbitrarily small times. On the other hand, if $F(u) = |u|^2 u$, the question of global existence of a solution with $L^2$ solution and the related question of (regular) wellposedness are widely open problems. Another important example of such critical problems is $d = 3$, $F(u) = |u|^4 u$, for which $s_c = 1$. In this case, Colliander–Keel–Staffilani–Takaoka–Tao [29] have recently proved global (regular) wellposedness.

**4.2. Operators on the real line.** The first kind of nonhomogeneous medium to which it is natural to generalize the above Strichartz inequalities – and hence the nonlinear results of Theorem 4.2 – is of course the real line, with an operator $P$ satisfying the assumptions (1.1), (1.2) of the introduction. In this case, the extension is rather straightforward for *local in time* Strichartz estimates. Indeed, if

$$P = -\frac{1}{\rho}\frac{d}{dx} a \frac{d}{dx},$$

the global change of variable $dy = (\rho(x)/a(x))^{1/2}\,dx$ and the conjugation $\lambda^{-1} P \lambda$ with the function $\lambda(y) = (a(x)\rho(x))^{-1/4}$ lead to the operator

$$\tilde{P} = -\frac{d^2}{dy^2} + V(y), \quad V = -\frac{\lambda''}{\lambda} + 2\left(\frac{\lambda'}{\lambda}\right)^2.$$

Since $V$ is a bounded function, the Strichartz inequalities for $\tilde{P}$ on a *finite time* interval $[0, T]$ with $C = C(T)$ are a straightforward consequence of the Euclidean ones, considering the term $Vu$ as a source term in the right hand side. We refer to [49] for a slightly different proof.

Though we shall not pursue in this direction, let us indicate that the question of singular coefficients $\rho, a$ has been addressed quite recently. In [2], Banica observed that, if $\rho = 1$ and $a$ is a piecewise constant function with a finite number of discontinuities, the dispersion estimate – and hence Strichartz inequalities – is valid, while it fails for an infinite number of discontinuities. Burq and Planchon [21] generalized this observation by proving that *global* Strichartz estimates hold as soon as $a$ has bounded variation. Notice that the BV regularity seems to be a relevant threshold, since, for every $s < 1$, $W^{s,1}$ functions $a$ are constructed in the appendix of [21] such that every kind of Strichartz inequality fails (see also an earlier result by Castro and Zuazua [22], which shows the same phenomenon for Hölder continuous functions $a$ of any exponent $\alpha < 1$).

**4.3. Strichartz inequalities with loss of derivative.** At this stage a natural question is of course to extend Strichartz inequalities (4.4) to variable coefficients in several space dimensions, as we did on the line. The following example shows that the situation is much more complicated. Our starting point is the following identity for the ground state of the harmonic oscillator,

$$(-h^2\partial_s^2 + s^2 - h)\big(\mathrm{e}^{-s^2/2h}\big) = 0.$$

Setting $s = r - 1$ and, for every positive integer $n$,

$$\psi_n(r, \theta) = \mathrm{e}^{-(r-1)^2/2h_n + in\theta}$$

where $h_n > 0$ is such that $h_n^{-2} = h_n^{-1} + n^2$, we infer

$$(-\partial_r^2 - \partial_\theta^2 - h_n^{-2}(1 - (r-1)^2))\psi_n = 0. \tag{4.5}$$

Notice that $\psi_n$ is the expression in polar coordinates of a smooth function on the complement of the origin in the plane, and that the operator

$$P_0 = -\frac{1}{1 - (r-1)^2}(\partial_r^2 + \partial_\theta^2)$$

is a positive elliptic operator of order 2 with smooth real coefficients, on the complement $\Omega$ of the origin in the disc of radius 2 endowed with the density $d\mu = (1 - (r-1)^2)\, dr\, d\theta = (2 - |x|)dx$. Since $P_0(1) = 0$, it follows that $P_0$ can be written $P_0 = -\frac{1}{\rho_0}\nabla.(A_0\nabla)$, where $\rho_0, A_0$ are smooth functions on $\Omega$ valued in positive numbers and definite positive matrices respectively. Notice that (4.5) reads $P_0\psi_n = h_n^{-2}\psi_n$, and that, as $n$ tends to $\infty$, $\psi_n$ is exponentially concentrating on the circle of radius 1. By cutting off $\psi_n$ near this circle, we obtain a sequence of

functions $\tilde{\psi}_n$ and a differential operator $P$ on $\mathbb{R}^2$ satisfying the assumptions of the introduction, such that

$$P\tilde{\psi}_n = h_n^{-2}\tilde{\psi}_n + r_n \tag{4.6}$$

where, for every $q \geq 1$ and for every $s \geq 0$,

$$\|\tilde{\psi}_n\|_{L^q} \sim n^{-1/2q}, \quad \|\tilde{\psi}_n\|_{H^s} \sim n^{s-1/4}, \quad \|r_n\|_{H^s} \leq C_s \mathrm{e}^{-\delta n}$$

for some $\delta > 0$. This sequence of functions is called a quasimode for the operator $P$. The geometric interpretation is that the circle of radius 1 is a (sufficiently stable) geodesic curve for the Riemannian metric defined by the principal symbol of $P$ (see e.g. Ralston [47]). By adding a suitable remainder term $w_n(t)$, we can write

$$u_n(t) := \mathrm{e}^{-itP}\tilde{\psi}_n = \mathrm{e}^{-ith_n^{-2}}\tilde{\psi}_n + w_n(t)$$

with $\|u_n\|_{L^p([0,T],L^q(\mathbb{R}^2))} \sim T^{1/p}\, n^{-1/2q}$ for every $p, q \geq 1$ and every $T > 0$. By using $u_n$ as a test function we conclude that, for every $q > 2$, for every $p \geq 1$, the estimate

$$\|u\|_{L^p([0,T],L^q)} \leq C\|u(0)\|_{L^2}$$

fails. More precisely, in view of the behavior of $u_n$ it is even impossible to replace the $L^2$ norm in the right hand side by the $H^s$ norm if

$$s < \frac{1}{4} - \frac{1}{2q}. \tag{4.7}$$

In other words, *in multidimensional heterogeneous media, losses of derivatives in Strichartz inequalities cannot be avoided.* The next question is of course to estimate precisely this loss of derivatives in terms of $p, q$ and the geometry of the medium. There is no need to say that this is a very difficult open problem. However, it is possible to give a general upper bound, which is valid for every geometry and already gives interesting applications to nonlinear problems.

**Theorem 4.3** (Staffilani–Tataru [53], Burq–Gérard–Tzvetkov [12]). *If $(p, q)$ is a d-admissible pair, the solution u of the equation*

$$i\partial_t u - Pu = f, \quad u(0) = u_0, \tag{4.8}$$

*satisfies the inequality*

$$\|u\|_{L^p([0,T],L^q(M))} \leq C_T \left(\|u_0\|_{H^{1/p}(M)} + \|f\|_{L^1([0,T],H^{1/p}(M))}\right). \tag{4.9}$$

*Proof* (*sketch*). Firstly, by Duhamel's formula and Minkowski's inequality, (4.9) is reduced to the case $f = 0$. Then, by Littlewood–Paley's analysis, we can assume that $u_0$ is spectrally supported in a dyadic interval, namely $\varphi(N^{-2}P)u_0 = u_0$ for some $\varphi \in C_0^\infty(\mathbb{R})$, where $N$ is a large dyadic integer. The advantage of this spectral

localization is that we can describe rather explicitly, by a standard semiclassical WKB analysis, the solution

$$u(t) = \mathrm{e}^{-itP}u_0$$

on a time interval of order $N^{-1} = h$. By a stationary phase argument, this implies the following dispersion estimate:

$$\|u(t)\|_{L^\infty(M)} \leq \frac{C}{|t|^{d/2}}\,\|u_0\|_{L^1(M)}, \quad |t| \lesssim \frac{1}{N}. \tag{4.10}$$

From this dispersion estimate, the $TT^*$ trick leads to the following semi-classical Strichartz inequalities

$$\|u\|_{L^p([0,N^{-1}],L^q(M))} \leq C\|u_0\|_{L^2(M)}, \quad u_0 = \varphi(N^{-2}P)u_0, \tag{4.11}$$

where $(p, q)$ stands for any $d$-admissible pair. The last step of the proof consists in iterating the estimates (4.11) on $N$ intervals of length $N^{-1}$ covering the interval $[0, 1]$. This yields a factor $N^{1/p}$ in the right hand side, and this completes he proof since

$$N^{1/p}\|u_0\|_{L^2(M)} \simeq \|u_0\|_{H^{1/p}(M)}. \qquad \square$$

**Remark 4.4.** Notice that when $d = 2$, the loss $\frac{1}{p} = \frac{1}{2} - \frac{1}{q}$ is twice as big as the threshold (4.7) derived from our counterexample in the beginning of this subsection. Indeed, the last step of the above proof may seem quite rough, since the decomposition of $[0, 1]$ into $N$ intervals of length $N^{-1}$ does not take into account the geometric features of $M$ and $P$. However, it is interesting to notice that there are geometries where some inequalities (4.9) are optimal. Indeed, if $M$ is compact and $d \geq 3$, inequality (4.9) with $p = 2$ applied to $f = 0$ and to the special Cauchy data $u_0 = \psi_\lambda$, where $\psi_\lambda$ is an eigenfunction of $P$ associated to a large eigenvalue $\lambda^2$, provides the estimate

$$\|\psi_\lambda\|_{L^q(M)} \leq C\,\lambda^{1/2}\|\psi_\lambda\|_{L^2(M)}, \quad q = \frac{2d}{d-2}. \tag{4.12}$$

Estimate (4.12) is one of the estimates obtained by Sogge [51] for the $L^r$ norms of the eigenfunctions of elliptic operators on compact manifolds, and it is known that this estimate is optimal if $M$ is the sphere $\mathbb{S}^d$ and $P$ is the standard Laplace operator, for spherical harmonics $\psi_\lambda$ which are functions of the distance to a fixed point (see [50]). A similar phenomenon occurs for $d = 2$ with $q = \infty$, except that our inequalities need an extra $\varepsilon$-derivative, due to the forbidden case $p = 2$, $q = \infty$.

Using inequalities (4.9), we obtain wellposedness results for nonlinear Schrödinger equations (1.3). For simplicity, we only state the case of polynomial nonlinearities.

**Corollary 4.5** (Burq–Gérard–Tzvetkov[12]). *Assume $d \geq 2$ and suppose that $F$ is a polynomial in $u$, $\bar{u}$ of degree $1 + \alpha \geq 2$. Then the Cauchy problem for* (1.3) *is regularly well-posed on $H^s(M)$ for*

$$s > \frac{d}{2} - \frac{1}{\max(\alpha, 2)}.$$

*Moreover, if $d = 3$ and $F(u) = |u|^2 u$, then the Cauchy problem for* (1.3) *is globally well-posed on $H^1(M)$, and it is globally regularly well-posed on $H^s(M)$ if $s > 1$.*

If $F$ is both gauge invariant and Hamiltonian with a nonnegative potential $V$, we can combine Corollary 4.5 with the conservation laws (1.9) and (1.7) to deduce global regular wellposedness on $H^1(M)$ if $d = 2$. In the special case of a cubic NLS in three space dimensions, observe that regular wellposedness is only known for $s > 1$; the uniform or regular wellposedness is still open in general on the energy space $H^1(M)$. This is in strong contrast with the case of the Euclidean case, where the critical nonlinearity is quintic. However, in Section 6 we shall improve Corollary 4.5 for several specific three-dimensional geometries.

We conclude this subsection by quoting a recent result concerning boundary value problems. In this case the WKB analysis is much more problematic, due to glancing rays. However it is possible to reduce the analysis, by a reflection argument, to the case of a boundaryless manifold endowed with a Lipschitz continuous Riemannian metric. Combining the method of proof of Theorem 4.3 with earlier smoothing ideas due to Bahouri and Chemin [3] (see also Tataru [57]) in the context of nonlinear wave equations, it is possible to obtain the following result, which, in the particular case of a plane domain, provides the first global wellposedness result for super-cubic nonlinearities (for the cubic case, earlier wellposedness results were due to Brezis–Gallouet [11] and Vladimirov [60], by different arguments).

**Theorem 4.6** (Anton [1]). *Assume that $M$ is a compact manifold and that $P$ is given by* (1.1) *where $\rho$ and $A$ are Lipschitz continuous. Then, for every $d$-admissible pair $(p, q)$, the solution of* (4.8) *satisfies*

$$\|u\|_{L^p([0,T],L^q(M))} \leq C_{s,T}\,(\|u_0\|_{H^s(M)} + \|f\|_{L^1([0,T],H^s(M))}), \quad s > \frac{3}{2p}. \tag{4.13}$$

*In particular, the estimate* (4.13) *still holds if $M$ is replaced by a smooth bounded open set in $\mathbb{R}^d$ if $-P$ is the Laplace operator $\Delta_D$ (resp. $\Delta_N$) with Dirichlet (resp. Neumann) boundary conditions and if the space $H^s(M)$ is replaced by the domain of the power $s/2$ of $P$. Consequently, if $d = 2$ and $F$ satisfies* (1.4)*, is gauge invariant and Hamiltonian with a nonnegative potential $V$, the equation* (1.3) *with Dirichlet (resp. Neumann) boundary condition has a unique global solution $u \in C(\mathbb{R}, H^1_0(\Omega))$ (resp. $u \in C(\mathbb{R}, H^1(\Omega))$) if $u_0 \in H^1_0(\Omega)$ (resp. $u_0 \in H^1(\Omega)$), and the map $u_0 \mapsto u$ is Lipschitz continuous.*

**4.4. Non-trapping metrics.** Though we are rather interested in new phenomena induced by the heterogeneity of the medium, we cannot conclude this section devoted to Strichartz inequalities without quoting a series of results giving sufficient conditions on the geometry of the operator $P$ on $M = \mathbb{R}^d$ in order that Euclidean inequalities (4.4) hold. All these conditions concern the Laplace operator with a non-trapping metric, namely a Riemannian metric on $\mathbb{R}^d$ such that no geodesic curve stays in a compact set during an arbitrarily long time: notice that this prevents counterexamples like

the one in the previous subsection. First of all, Staffilani–Tataru [53] proved (4.4) on finite time intervals if the non-trapping metric is the perturbation of the Euclidean metric by a $C^2$ compactly supported function. Then Robbiano–Zuily [48] generalized this result to short range perturbations of the Euclidean metric by a very precise parametrix construction. Similar results were obtained by Hassell–Tao–Wunsch on asymptotically conic manifolds, using different methods. Finally, Bouclet–Tzvetkov [5] recently tackled the case of long range perturbations of the Euclidean metric. Notice that the proofs in [53] and [5] rely on the local smoothing effect for non-trapping metrics (see Doi [30]) which appears to be the complementary property of estimates (4.9) to obtain Strichartz inequalitites (4.4) without loss on finite time intervals. Finally, we refer to [15] for applications of this smoothing effect for non-trapping exterior domains to boundary problems for nonlinear Schrödinger equations.

## 5. Multilinear Strichartz estimates

In Subsection 4.3, we observed that, in several space dimensions, the geometry of the medium may induce losses of derivatives in Strichartz inequalities, and that some of these losses are optimal in specific geometries such as the sphere. The wellposedness results deduced from these Strichartz inequalities with loss are altered with respect to the Euclidean case – compare Theorem 4.2 and Corollary 4.5. However, so far we did not give evidence of this alteration. Indeed, as we shall see in the sequel, the whole range of Strichartz inequalities is not necessary to give optimal wellposedness results for NLSs. In order to understand this, we begin with revisiting the question of regular wellposedness for the cubic NLS.

**5.1. A criterion for regular wellposedness of the cubic NLS.** Given a dyadic integer $N$, let us say that a function $u$ on $M$ is *spectrally localized at frequency $N$* if

$$\mathbf{1}_{[N,2N]}(\sqrt{1+P})(u) = u. \tag{5.1}$$

We start with the important notion of bilinear Strichartz estimate, which originates in the works of Bourgain [6] and of Klainerman–Machedon [45] in the context of null forms for the wave equation.

**Definition 5.1.** Let $s \geq 0$. We shall say that *the Schrödinger group for $P$ satisfies a bilinear Strichartz estimate of order $s$ on $M$* if there exists a constant $C$ such that, for all dyadic integers $N$, $L$, for all functions $u_0$, $v_0$ on $M$ spectrally localized at frequencies $N$, $L$ respectively, the functions

$$u(t) = \mathrm{e}^{-itP}(u_0), \quad v(t) = \mathrm{e}^{-itP}(v_0)$$

satisfy the inequality

$$\|uv\|_{L^2([0,1]\times M)} \leq C \min(N, L)^s \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}. \tag{5.2}$$

Notice that by setting $v_0 = u_0$, $L = N$ and by using the Littlewood–Paley inequality, one easily shows that a bilinear Strichartz estimate of order $s$ implies a Strichartz-type estimate of the space-time $L^4$ norm of a solution to the linear Schrödinger equation in terms of the $H^{s/2}$ norm of the Cauchy data. However, if $s > 0$, a bilinear Strichartz estimate says more, since the price to pay for estimating the $L^2$ norm of a product of such solutions only involves the lowest frequency of these solutions. The importance of bilinear Strichartz estimates in the wellposedness theory for NLSs clearly appears in the following theorem, which is a slight reformulation of some results from [16].

**Theorem 5.2** (Burq–Gérard–Tzvetkov [16]). *Assume $F(u) = \pm|u|^2u$ and $s \geq 0$.*

i) *If the Cauchy problem for* (1.3) *is regularly well-posed on $H^s(M)$, then the Schrödinger group for $P$ satisfies a bilinear Strichartz estimate of order $s$ on $M$.*

ii) *If the Schrödinger group for $P$ satisfies a bilinear Strichartz estimate of order $s$ on $M$, then, for every $\sigma > s$, the Cauchy problem for* (1.3) *is regularly well-posed on $H^\sigma(M)$.*

In other words, the existence of a bilinear Strichartz estimate of order $s$ is *almost* a criterion for regular wellposedness for cubic NLSs on $H^s(M)$. Notice that the strict inequality $\sigma > s$ in ii) cannot be extended to an equality in general. Indeed, for the Euclidean Laplace operator on $\mathbb{R}^2$, the Strichartz inequality (4.4) for the admissible pair (4, 4) combined with the Hölder inequality implies a bilinear estimate of order 0; however, we already observed at the end of Subsection 4.1 that the focusing cubic equation with $F(u) = -|u|^2u$ is not well-posed on $L^2(\mathbb{R}^2)$.

Another comment on the above statement is that it provides counterexamples to regular wellposedness. Indeed, in the beginning of Subsection 4.3 we constructed an example of an operator $P$ on $\mathbb{R}^2$ such that the estimate

$$\|u\|_{L^4([0,1]\times\mathbb{R}^2)} \leq C\|u_0\|_{H^s(\mathbb{R}^2)}$$

fails if $s < 1/8$. Consequently, the Schrödinger group for this operator does not enjoy bilinear Strichartz estimates of order $s < 1/4$, and thus *the cubic NLS for this operator is not regularly well-posed on $H^s(\mathbb{R}^2)$ for $s < 1/4$*. This is in strong contrast with the Euclidean case, where we know that regular wellposedness holds for every $s > 0$ (see Theorem 4.2).

*Proof* (*sketch*). The proofs of parts i) and ii) of Theorem 5.2 are of unequal length and difficulty. As a matter of fact, using regular wellposedness of (1.3) for smooth data stated in Proposition 3.1 and the propagation of regularity contained in the definition of regular wellposedness, it is easy to check that the third differential at 0 of the map $\Phi_1 : u_0 \mapsto u(1)$ is given by the following polarized form of the first iteration of the Duhamel equation (3.1):

$$D^3\Phi_1(0)(u_0, u_0, v_0) = -2i \int_0^1 \mathrm{e}^{-i(1-t)P}(2|u(t)|^2v(t) + u(t)^2\overline{v(t)})\,dt.$$

Here we assumed $N \leq L$ without loss of generality. We now compute the scalar product of both members of the above identity with $\mathrm{e}^{-iP}v_0$, and we use the assumed continuity of the trilinear map $D^3\Phi_1(0)$ from $(H^s)^3$ to $H^s$. This yields

$$\|uv\|^2_{L^2([0,1]\times M)} \leq C\|u_0\|^2_{H^s}\,\|v_0\|_{H^s}\,\|v_0\|_{H^{-s}}.$$

Using that

$$\|f\|_{H^{\pm s}} \simeq N^{\pm s}\|f\|_{L^2}$$

if $f$ is spectrally supported at frequency $N$, we infer the bilinear Strichartz estimate (5.2), and hence part i) is proved.

Let us come to part ii). The main idea, which in this context is due to Bourgain [6], is to introduce the scale of Hilbert spaces

$$X^{s,b}(P, \mathbb{R}\times M) = \{v \in \mathcal{S}'(\mathbb{R}\times M) : (1+|i\partial_t - P|^2)^{b/2}\,(1+P)^{s/2}v \in L^2(\mathbb{R}\times M)\}$$

for $s, b \in \mathbb{R}$. We refer to [35] for a pedagogical introduction to this strategy. Denoting by $X^{s,b}_T(P)$ the space of restrictions of elements of $X^{s,b}(P, \mathbb{R}\times M)$ to $]-T, T[\times M$, it is easy to observe that

$$X^{s,b}_T(P) \subset C([-T, T], H^s(M)) \quad \text{for all } b > \tfrac{1}{2},$$

and that the solution of the linear Schrödinger equation with datum in $H^s$ lies in $X^{s,b}_T(P)$ for every $b$. Moreover, the Duhamel term in the integral equation (3.1) can be handled by means of these spaces as

$$\left\| \int_0^t \mathrm{e}^{-i(t-t')P} f(t')\,dt' \right\|_{X^{s,b}_T(P)} \leq C\,T^{1-b-b'}\|f\|_{X^{s,-b'}_T(P)}$$

if $0 < T \leq 1$, $0 < b' < \frac{1}{2} < b$, $b + b' < 1$. The crux of the proof is then to observe that a bilinear Strichartz estimate of order $s$ implies the following estimates, for $\sigma' \geq \sigma > s$ and suitable $b, b'$ as above,

$$\|v_1\bar{v}_2v_3\|_{X^{\sigma,-b'}(P)} \leq C\,\|v_1\|_{X^{\sigma,b}(P)}\,\|v_2\|_{X^{\sigma,b}(P)}\|v_3\|_{X^{\sigma,b}(P)},$$
$$\||v|^2v\|_{X^{\sigma',-b'}(P)} \leq C\,\|v\|^2_{X^{\sigma,b}(P)}\,\|v\|_{X^{\sigma',b}(P)},$$

which allow the use a fixed point argument in $X^{\sigma,b}_T(P)$ in the resolution of the integral equation (3.1). □

**Remark 5.3.** Combining the Strichartz inequalities of Theorem 4.3 with the Sobolev inequalities, one easily shows that, if $d \geq 2$ and if $u_0$ spectrally localized at frequency $N$, the solution of the linear Schrödinger equation satisfies

$$\|u\|_{L^2([0,1],L^\infty(M))} \leq C_s N^s \|u_0\|_{L^2(M)} \quad \text{for all } s > \tfrac{d-1}{2}.$$

Using Hölder's inequality and the conservation of the $L^2$ norm by the Schrödinger group, we infer, in the context of Definition 5.1,

$$\begin{aligned}\|uv\|_{L^2([0,1]\times M)} &\leq \|u\|_{L^2([0,1],L^\infty(M))}\,\|v\|_{L^\infty([0,1],L^2(M))}\\ &\leq C_s N^s\,\|u_0\|_{L^2(M)}\|v_0\|_{L^2(M)},\end{aligned}$$

namely a bilinear Strichartz estimate of order $s > (d-1)/2$. Applying Theorem 5.2, we conclude that a cubic NLS is regularly well-posed on $H^s(M)$ for every $s > (d-1)/2$ if $d \geq 2$, which is consistent with Corollary 4.5. However, we shall see in the next section that this bilinear Strichartz estimate is far from optimal in several specific cases, therefore the threshold of regular wellposedness will be improved through Theorem 5.2.

**5.2. Generalization to subcubic nonlinearities.** Bilinear Strichartz estimates can also be used to prove uniform wellposedness for (1.3) when $F$ is not polynomial. In particular, combining the method of proof of part ii) in Theorem 5.2 with paradifferential expansions, it is possible to prove the following result.

**Theorem 5.4.** *Assume that the Schrödinger group for $P$ satisfies a bilinear Strichartz estimate of order $s$ on $M$ and that $F$ satisfies* (1.4) *with $\alpha \leq 2$. Then the Cauchy problem for* (1.3) *is uniformly well-posed on $H^\sigma(M)$ for every $\sigma > s$.*

Compared to Theorem 4.2 it may seem surprising that the regularity threshold of uniform wellposedness for (1.3) does not depend on $\alpha$. However we shall see such an example in Section 6, in the case of the two-dimensional sphere.

**5.3. Higher order nonlinearities.** By mimicking Definition 5.1 it is easy to define the notion of $k$-linear estimate for $k \geq 3$.

**Definition 5.5.** Let $k$ be an integer $\geq 3$ and $s_1, \ldots, s_{k-1} \geq 0$ . We shall say that *the Schrödinger group for $P$ satisfies a $k$-linear Strichartz estimate of order* $(s_1, \ldots, s_{k-1})$ *on $M$* if there exists a constant $C$ such that, for all dyadic integers $N_1 \leq \cdots \leq N_k$, for all functions $u_{1,0}, \ldots, u_{k,0}$ on $M$ spectrally localized at frequencies $N_1, \ldots, N_k$ respectively, the functions

$$u_j(t) = \mathrm{e}^{-itP}(u_{j,0}), \quad j = 1, \ldots, k,$$

satisfy the inequality

$$\|u_1 \ldots u_k\|_{L^2([0,1]\times M)} \leq C\, N_1^{s_1} \ldots N_{k-1}^{s_{k-1}} \|u_{1,0}\|_{L^2(M)} \ldots |u_{k,0}\|_{L^2(M)}. \tag{5.3}$$

**Remark 5.6.** By an iterated use of Hölder's inequality, we can always assume $s_1 \geq \cdots \geq s_{k-1}$.

Next we have the equivalent of Theorem 5.2.

**Theorem 5.7.** *Let $s \geq 0$, $m$ be an integer $\geq 2$ and $F(u) = \pm|u|^{2m}u$. Then*

i) *If* (1.3) *is regularly well-posed on $H^s(M)$, the Schrödinger group for $P$ satisfies an $(m+1)$-linear Strichartz estimate of order $(s, \dots, s)$.*

ii) *If the Schrödinger group for $P$ satisfies an $(m+1)$-linear Strichartz estimate of order $(s, \dots, s)$, then* (1.3) *is regularly well-posed on $H^\sigma(M)$ for every $\sigma > s$.*

Moreover, the use of different exponents in the list $s_1, \dots, s_k$ can help to tackle non-polynomial nonlinearities. Let us give an example for nonlinearities which are intermediate between cubic and quintic, which is essentially borrowed from [17].

**Theorem 5.8.** *Assume the Schrödinger group for $P$ satisfies a trilinear Strichartz estimate of order $(s_1, s_2)$ with $s_1 > s_2 \geq 0$ and $M$ is compact. Let $F$ satisfy* (1.4) *with $2 < \alpha < 4$. Then* (1.3) *is uniformly well-posed on $H^s(M)$ for every*

$$s > \left(1 - \frac{2}{\alpha}\right) s_1 + \frac{2}{\alpha} s_2.$$

**5.4. Multilinear estimates for spectral projectors.** If $M$ is compact, a $k$-linear Strichartz estimate for the Schrödinger implies a $k$-linear estimate for eigenfunctions of $P$ of the following kind:

$$\|\varphi_1 \dots \varphi_k\|_{L^2(M)} \leq C\, \lambda_1^{s_1} \dots \lambda_{k-1}^{s_{k-1}} \|\varphi_1\|_{L^2(M)} \dots \|\varphi_k\|_{L^2(M)},$$

$\varphi_j$ an eigenfunction of $P$ associated to the eigenvalue $\lambda_j^2$ and $1 \leq \lambda_1 \leq \dots \leq \lambda_k$. These estimates can be seen as $k$-linear versions of Sogge's estimates [51], [52]. A first step is therefore to decide for which orders such $k$-linear estimates hold. The following result gives a fairly general answer to this question. For the sake of generality, we deal, as in [51], [52], with spectral projectors $\Pi_\lambda = \mathbf{1}_{\lambda \leq \sqrt{P} \leq \lambda+1}$ on clusters of bounded length for the square root of $P$. Notice that this is a much more stringent spectral localization than the dyadic one which we introduced in the beginning of this section. Under this form our result makes sense in the case $M = \mathbb{R}^d$ too.

**Theorem 5.9** (Burq–Gérard–Tzvetkov [17]). *We have the bilinear estimates*

$$\|\Pi_{\lambda_1} f_1\, \Pi_{\lambda_2} f_2\|_{L^2(M)} \leq C\, \|f_1\|_{L^2(M)} \|f_2\|_{L^2(M)} \begin{cases} \lambda_1^{1/4} & \text{if } d = 2, \\ (\lambda_1\, \log(\lambda_1))^{1/2} & \text{if } d = 3, \\ \lambda_1^{(d-2)/2} & \text{if } d \geq 4 \end{cases}$$

*if $2 \leq \lambda_1 \leq \lambda_2$. In the special case $d = 2$, we have the trilinear estimate*

$$\|\Pi_{\lambda_1} f_1\, \Pi_{\lambda_2} f_2\, \Pi_{\lambda_3} f_3\|_{L^2(M)} \leq C(\lambda_1\lambda_2)^{1/4} \|f_1\|_{L^2(M)}\, \|f_2\|_{L^2(M)}\, \|f_3\|_{L^2(M)}$$

*if $1 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3$.*

**Remark 5.10.** The logarithmic factor in the right hand side of the bilinear estimate in three space dimensions may be just technical. Apart from this, the linear estimates deduced from the above bilinear and trilinear estimates by making all the frequencies $\lambda_j$ equal and all the functions $f_j$ equal, are exactly the $L^4$ and $L^6$ estimates among the $L^p$ estimates proved by Sogge in [51], which are known to be optimal. Moreover, all the exponents in the multilinear estimates of Theorem 5.9 are optimal in the particular case of the sphere.

Finally, we did not state the other multilinear estimates for those spectral projectors, since they are essentially straightforward consequences of the above ones and of the $L^\infty$ estimate due to Sogge,

$$\|\Pi_\lambda f\|_{L^\infty(M)} \leq C\lambda^{\frac{d-1}{2}}\|f\|_{L^2(M)}, \quad \lambda \geq 1. \tag{5.4}$$

*Proof* (*sketch*). There are several proofs of this result (see [16] for the bilinear estimate in the two-dimensional case, and [17] or [18] for the general case). Here we follow [18]. We set $u = \Pi_\lambda f$, $h = \lambda^{-1}$, and we observe that the spectral localization given by $\Pi_\lambda$ can be formulated as a semiclassical PDE:

$$h^2 Pu - u = h\, r, \quad \|r\|_{L^2(M)} \leq C\|u\|_{L^2(M)}. \tag{5.5}$$

The main observation is then that this equation can be seen microlocally as an evolution equation with respect to one of the spatial coordinates, say $x_1$. For this equation it is possible to perform exactly the same analysis as we did in the proof of Theorem 4.3. This yields semiclassical versions of Strichartz inequalities (4.9) for $(d-1)$-admissible pairs, and the result follows by using Hölder's inequality as in Remark 5.3. □

## 6. The case of some simple compact manifolds

In this section we investigate the Cauchy problem for (1.3) on tori, spheres and balls, where $-P$ is the standard Laplace operator, trying to improve the results of Section 4. The basic tools for positive results of wellposedness are borrowed from Section 5. Several illposedness results are also obtained by explicit constructions. Moreover, we point some open problems in this context.

**6.1. Tori.** The first compact manifolds on which the global Cauchy problem for (1.3) has been studied are the tori

$$\mathbb{T}^d = \mathbb{R}^d/(2\pi\mathbb{Z})^d.$$

The main reference is the fundamental work of Bourgain [6], [7] (see also [8] and Ginibre [35]). We start with the one-dimensional case.

**Theorem 6.1** (Bourgain [6]). *If* $F(u) = \pm|u|^2u$, *then* (1.3) *is globally regularly well-posed on* $L^2(\mathbb{T})$.

*If $F(u) = \pm|u|^4 u$, then* (1.3) *is locally regularly well-posed on $H^\varepsilon(\mathbb{T})$, but not on $L^2(\mathbb{T})$.*

If we compare with Theorem 4.2, we see that the quintic nonlinearity is critical for $L^2$ wellposedness on the line. However, on the line a quintic NLS is regularly well-posed on small neighborhoods of 0 in $L^2(\mathbb{R})$, while it is not the case on the circle.

The proof of Theorem 6.1 essentially combines Theorem 5.7 with the explicit spectral representation for the solution of the linear Schrödinger equation,

$$u(t,x) = \mathrm{e}^{it\partial_x^2} u_0\,(x) = \sum_{n\in\mathbb{Z}} \mathrm{e}^{-itn^2} \mathrm{e}^{inx} \hat{u}_0(n),$$

which yields the needed bilinear and trilinear estimates by a direct calculation based on the Parseval formula in both time and space variables, following an old idea due to Zygmund [63]. The lack of regular wellposedness of a quintic NLS on $L^2$ results from the first part of Theorem 5.7 and an explicit example of a sequence of data $u_0$ such that $\|u\|_{L^6(\mathbb{T}^2)}\,\|u_0\|^{-1}_{L^2(\mathbb{T})}$ is not bounded.

Let us say a word about data in $H^s(\mathbb{T})$ for $s < 0$. In this case, there are several results of illposedness for (1.3). The most elementary one (see [13]) is the lack of uniform wellposedness which can be deduced from the explicit solutions

$$u(t,x) = \kappa n^{|s|} \mathrm{e}^{-it(n^2 + G'(\kappa^2 n^{2|s|}))} \mathrm{e}^{inx}$$

if $F(z) = G'(|z|^2)z$. Notice that, if $G'$ behaves like a power at infinity, a small variation of $\kappa$ around 1 induces a large variation of the phase shift of $u(t,x)$ for $t > 0$ and $n$ large, whence some lack of stability. Notice that a similar argument had been used earlier by Birnir–Kenig–Ponce–Svanstedt–Vega [4] and by Kenig–Ponce–Vega [44] on the line. Furthermore, by a careful study of the interaction between high and low frequencies, Christ, Colliander and Tao prove in [28] that, for the cubic case, the flow map fails to be continuous from any ball of $H^s$ to the space of distributions. As for the quintic equation, the flow map fails to be uniformly continuous from any ball of $H^s$ endowed with the topology of $C^\infty$, to the space of distributions. More recently, Christ [25] constructed non trivial weak solutions in $C_t(H^s)$ of a modified version of a cubic NLS, with zero Cauchy datum.

Let us come to the multidimensional case. Using the same method as for Theorem 6.1, it is possible to prove multilinear Strichartz estimates on $\mathbb{T}^d$ for $d \geq 2$. The interesting point is that these estimates are the same as on $\mathbb{R}^d$, except the fact that they are local in time, and that a loss $\varepsilon$ may alter the orders. For instance, the Schrödinger group on $\mathbb{T}^d$ enjoys a bilinear Strichartz estimate of order $(d-2)/2 + \varepsilon$ for every $\varepsilon > 0$. In view of Theorems 5.2, 5.7 and 5.8, this implies in particular the following results.

**Theorem 6.2** (Bourgain [6]). *If $F(u) = \pm|u|^2 u$, then* (1.3) *is regularly well-posed on $H^\varepsilon(\mathbb{T}^2)$ for every $\varepsilon > 0$.*

*If $F$ satisfies* (1.4) *for some $\alpha \in [2,4]$, then* (1.3) *is uniformly wellposed on $H^s(\mathbb{T}^3)$ for every $s > 3/2 - 2/\alpha$. In particular, if moreover $\alpha < 4$, $F$ is gauge invariant and Hamiltonian with a nonnegative potential,* (1.3) *is globally uniformly well-posed on $H^1(\mathbb{T}^3)$.*

In the case $d = 4$, the regularity $s = 1$ is critical for the cubic NLS. However, by means of logarithmic estimates based on a careful study of exponential sums, global wellposedness in $H^s(\mathbb{T}^4)$, $s > 1$, can be obtained for (1.3) with nonlinearities with quadratic growth such as $F = F_{1,+}$ (see Bourgain [7]).

Let us conclude this subsection by quoting two open problems. The first one concerns the quintic defocusing problem, namely $F(u) = |u|^4 u$, on $\mathbb{T}^3$. According to Theorem 6.2, it is regularly wellposed on $H^s(\mathbb{T}^3)$ for every $s > 1$, but nothing is known about $s = 1$, even for small data. This would yield global regular wellposedness in view of conservation laws.

The second open problem is the generalization of the above results to tori of the type

$$\mathbb{T}^d(\theta_1, \dots, \theta_d) = \mathbb{R}/\theta_1\mathbb{Z} \times \cdots \times \mathbb{R}/\theta_d\mathbb{Z},$$

where the $\theta_j$'s are positive numbers, possibly irrationally independent. The possibly chaotic behavior of the spectrum

$$\lambda = \omega_1^2 n_1^2 + \cdots + \omega_d^2 n_d^2, \quad \omega_j = \frac{2\pi}{\theta_j},\ n_j \in \mathbb{Z},$$

makes the multilinear Strichartz estimates particularly delicate to obtain. For instance, if $d = 2$, the optimal order of the bilinear Strichartz estimate – and thus the threshold of regularity wellposedness of cubic NLSs – is not known. However, if $d = 3$, it is possible to prove that the local $L_t^p(L_x^4)$ norm of the solution of the linear Schrödinger equation scales as on the Euclidean space, for any $p > 16/3$ (see Bourgain [9]). This implies a trilinear estimate of order $\left(\frac{5}{4} - \varepsilon, \frac{3}{4} + \varepsilon\right)$ for every $\varepsilon > 0$, and, by Theorem 5.8, global uniform wellposedness of the defocusing subquintic NLS on $H^1(\mathbb{T}^3)$. Moreover, by more refined counting arguments, it is possible to reduce the order of the bilinear Strichartz estimate from $\frac{3}{4} + \varepsilon$ to $\frac{2}{3} + \varepsilon$ ([9]).

**6.2. Spheres.** The case of multidimensional spheres is of course very natural, since we observed in Subsection 4.3 that the loss in endpoint Strichartz inequalities is optimal on them. Therefore we could expect that the wellposedness results for (1.3) are the worst ones on spheres. The two-dimensional case is particularly interesting in this respect. The following theorem is a slight generalization of results in [13] and [16].

**Theorem 6.3** (Burq–Gérard–Tzvetkov [13], [16]). *The Cauchy problem for the cubic NLS, i.e.* (1.3) *with $F(u) = \pm|u|^2 u$, is regularly well-posed on $H^s(\mathbb{S}^2)$ for every $s > 1/4$ and* not *uniformly well-posed on $H^s(\mathbb{S}^2)$ for every $s < 1/4$.*

*If $F = F_{\alpha,\pm}$ (see* (1.10)*) with $\alpha \in ]0, 2]$, it is uniformly well-posed on $H^s(\mathbb{S}^2)$ for every $s > 1/4$, and* not *uniformly well-posed for $s < 1/4$.*

*The Cauchy problem for the quintic NLS, i.e.* (1.3) *with $F(u) = \pm|u|^4u$ is uniformly well-posed on $H^s(\mathbb{S}^2)$ for every $s > 1/2$.*

**Remark 6.4.** The first striking fact contained in the above result is that, unlike cubic NLSs on the Euclidean plane or on the (square) torus, a cubic NLS on the sphere has a threshold of regular (or uniform) wellposedness which is $1/4$ and not 0. Notice that we already met this exponent $1/4$ in Subsection 5.1, in connection with the counterexample of Subsection 4.3. In fact, the geometric phenomenon is the same here, namely concentration on a closed geodesic, but specific information about the sphere allow to get optimal results. A very natural open problem is of course to decide whether the *wellposedness* threshold is also $1/4$, or if it is smaller (for instance 0), as we quoted in Section 2 about the modified KdV equation.

Another important open question is raised by the comparison of the above result with Corollary 4.5, which, for a cubic NLS on general surfaces, needs $s > 1/2$ for regular wellposedness. In fact, we ignore if there is a compact surface where a cubic NLS is not regularly well-posed on $H^s$ for some $s \in ]1/4, 1/2]$.

A third observation is that, unlike positive thresholds on the Euclidean space, the one on the sphere is not always changing with the parameter $\alpha$. Indeed, for $0 < \alpha < 2$, it is frozen at $1/4$.

Finally, for the quintic NLS, the last statement of Theorem 6.3, combined with Theorem 3.3 shows that the three wellposedness thresholds coincide with $1/2$, which is the Euclidean one. This suggests a general mechanism which we already met in the context of tori, namely that *the $L^p_t(L^q_x)$ estimates of the solutions of the linear Schrödinger equation seem to become as good as the Euclidean ones if $p, q$ are large enough, so that, for $\alpha$ large enough, the wellposedness thresholds become identical to the Euclidean ones.* We shall check this phenomenon for higher dimensional spheres as well. However, we do not have any argument for proving it on a general compact manifold.

*Proof* (*sketch*). The positive results on uniform and regular wellposedness are consequences of Theorems 5.2, 5.4 and 5.7 and of multilinear Strichartz estimates for the Schrödinger group on $\mathbb{S}^2$. As in the previous subsection, we use the exact spectral representation of solutions to the linear Schrödinger equation,

$$u(t, x) = \sum_{n\sim N} \mathrm{e}^{-itn(n+1)} H_n(x),$$

where $H_n$ are spherical harmonics of degree $n$, and the condition $n \sim N$ corresponds to the spectral localization at frequency $N$. Using Parseval formula in the time variable and multilinear spectral estimates given by Theorem 5.9, we obtain a bilinear Strichartz estimate of order $1/4+\varepsilon$ and a trilinear estimate of order $(3/4+\varepsilon, 1/4+\varepsilon)$ for every $\varepsilon > 0$. The latter is even better than what we need. In particular, using Theorem 5.8, we infer that (1.3), with $F = F_{\alpha,\pm}$ (see (1.10)) and $\alpha \in ]2, 4[$, is uniformly

well-posed on $H^s(\mathbb{S}^2)$ for every $s > 3/4 - 1/\alpha$. However we do not know if this threshold is optimal.

Here an important role is played by the localization of the spectrum around squares of the integers, so that this proof can only be generalized to Zoll manifolds (see [12] and [16] for details).

We come now to the illposedness result. We observe that, as in the counterexample given in Subsection 4.3, the following sequence of spherical harmonics

$$\psi_n(x) = (x_1 + ix_2)^n, \quad x_1^2 + x_2^2 + x_3^2 = 1$$

is concentrating exponentially on the closed geodesic $x_1^2 + x_2^2 = 1$ and satisfies

$$\|\psi_n\|_{L^q} \simeq n^{\frac{1}{4} - \frac{1}{2q}} \|\psi_n\|_{L^2}, \quad q \geq 2.$$

Moreover, $\psi_n$ has the remarkable property that it is the ground state of the Laplace operator on the space $V_n$ of functions $f$ satisfying the symmetry property

$$f(R_\theta(x)) = \mathrm{e}^{in\theta} f(x),$$

where $R_\theta$ is the rotation of angle $\theta$ around the $x_3$ axis. The idea is to construct stationary solutions to (1.3) by minimizing the energy $H(f)$ on the $L^2$ sphere of $V_n$ for small radii $\delta_n = \kappa_n n^{-s}$, for different values of the parameter $\kappa_n \sim 1$. It turns out that the minimizers $f_n$ are very close to the line directed by $\psi_n$, and that the nonlinear eigenvalue $\omega_n$ can be precisely estimated as $n$ goes to $\infty$, creating for the solution

$$u_n(t, x) = \mathrm{e}^{-it\omega_n} f_n(x)$$

the same kind of instability that we already observed in the case of the one-dimensional torus. We refer to [31] for details, or to [13] for a slightly different approach. □

Finally, we observe that the above methods can be applied to higher-dimensional spheres. We gather the most striking facts in the following theorem.

**Theorem 6.5** (Burq–Gérard–Tzvetkov). *The Cauchy problem for a cubic NLS is regularly well-posed on $H^s(\mathbb{S}^3)$ for $s > 1/2$, and not uniformly well-posed for $s < 1/2$* ([17], [13]).

*If $F = F_{\alpha,\pm}$ (see (1.10)), it is uniformly well-posed on $H^s(\mathbb{S}^3)$ if $s > s(\alpha)$, and not uniformly well-posed for $s < s(\alpha)$, with $s(\alpha) = 1/2$ if $\alpha \leq 2$, and $s(\alpha) = 3/2 - 2\alpha$ if $\alpha \in [2, 4]$. In particularly it is globally uniformly well-posed on $H^1(\mathbb{S}^3)$ if $\alpha < 4$* ([17], [13]).

*The Cauchy problem for a cubic NLS is regularly well-posed on $H^s(\mathbb{S}^4)$ if $s > 1$, but not for $s = 1$* ([16], [12]).

*The Cauchy problem for* (1.3) *with $F = F_{\alpha,\pm}$ is not uniformly well-posed on $H^1(\mathbb{S}^6)$ for every $\alpha \in ]0, 1]$* ([13]).

In the case $d = 3$, as in the torus case, the question of (regular) wellposedness of the quintic NLS on small data in $H^1$ remains open. See however some partial results in this direction in [19]. As for global wellposedness of the subquintic case, it is also known on $\mathbb{S}^2 \times \mathbb{T}$, but is completely open for an arbitrary three–manifold.

In the case $d = 4$, global wellposedness for some smoothed variants of the cubic NLS can be found in [32].

Finally, let us emphasize that the illposedness result on $\mathbb{S}^6$ is in strong contrast with the case $d = 6$ in Theorem 4.2.

**6.3. Balls.** If $-P$ is the Laplace operator on the ball $\mathbb{B}^d$ of the $d$-dimensional Euclidean space with Dirichlet or Neumann boundary conditions, it is possible to take advantage of stronger concentration phenomena of the eigenfunctions at the boundary to produce illposedness for higher regularity.

**Theorem 6.6** (Burq–Gérard–Tzvetkov [14], [20]). *Let $-P$ be the Laplace operator on $\mathbb{B}^d$ with Dirichlet (resp. Neumann) boundary conditions.*

*If $d = 2$, the Cauchy problem for the cubic NLS is not uniformly well-posed on the domain of $P^{s/2}$ for $s \in [0, 1/3[$.*

*The Cauchy problem for* (1.3) *with $F = F_{\alpha,\pm}$ (see* (1.10)) *is not uniformly well-posed on $H^1_0(\mathbb{B}^5)$ (resp. $H^1(\mathbb{B}^5)$) for every $\alpha \in ]0, 1]$.*

Finally, let us mention that, contrarily to the case of spheres, global wellposedness for subquintic (1.3) with boundary conditions on $\mathbb{B}^3$ remains an open problem.

## References


[1] Anton, R., Strichartz inequalities for Lipschitz metrics on manifolds and nonlinear Schrödinger equations on domains. Preprint, January 2006.

[2] Banica, V., Dispersion and Strichartz inequalities for Schrödinger equations with singular coefficients. *SIAM J. Math. Anal.* **35** (2003), 868–883.

[3] Bahouri, H., Chemin, J.-Y., Equations d'ondes quasilinéaires et estimations de Strichartz. *Amer. J. Math.* **121** (1999), 1337–1377.

[4] Birnir, B., Kenig, C., Ponce, G., Svanstedt, N., Vega, L., On the ill-posedness of the IVP for the generalized KdV and nonlinear Schrödinger equation. *J. London Math. Soc.* **53** (1996), 551–559.

[5] Bouclet, J.-M., Tzvetkov, N., Strichartz estimates for long range perturbations. Preprint, September 2005.

[6] Bourgain, J., Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations I. Schrödinger equations. *Geom. Funct. Anal.* **3** (1993), 107–156.

[7] Bourgain, J., Exponential sums and nonlinear Schrödinger equations. *Geom. Funct. Anal.* **3** (1993) 157–178.

[8] Bourgain, J., *Global Solutions of Nonlinear Schrödinger equations.* Amer. Math. Soc. Colloq. Publ. 46, Amer. Math. Soc., Providence, RI, 1999.

[9] Bourgain, J., Remarks on Strichartz' inequalities on irrational tori. Preprint, 2004; in *Mathematical Aspects of nonlinear PDE*, Ann. of Math. Studies, Princeton University Press, Princeton, NJ, to appear.

[10] Bourgain, J., Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations II. The KdV equation. *Geom. Funct. Anal.* **3** (1993), 157–178.

[11] Brezis, H., Gallouet, T., Nonlinear Schrödinger evolution equations. *Nonlinear Anal.* **4** (1980), 677–681.

[12] Burq, N., Gérard, P., Tzvetkov, N., Strichartz inequalities and the nonlinear Schrödinger equation on compact manifolds. *Amer. J. Math.* **126** (2004), 569–605.

[13] Burq, N., Gérard, P., Tzvetkov, N., An instability property of the nonlinear Schrödinger equation on $S^d$. *Math. Res. Lett.* **9** (2002), 323–335.

[14] Burq, N., Gérard, P., Tzvetkov, N., Two singular dynamics of the nonlinear Schrödinger equation on a plane domain. *Geom. Funct. Anal.* **13** (2003), 1–19.

[15] Burq, N., Gérard, P., Tzvetkov, N., On nonlinear Schrödinger equations in exterior domains. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **21** (2004), 295–318.

[16] Burq, N., Gérard, P., Tzvetkov, N., Bilinear eigenfunction estimates and the nonlinear Schrödinger equation on surfaces. *Invent. Math.* **159** (2005), 187–223.

[17] Burq, N., Gérard, P., Tzvetkov, N., Multilinear eigenfunction estimates and global existence for the three dimensional nonlinear Schrödinger equations. *Ann. Sci. École Norm. Sup.* **38** (2005), 255–301.

[18] Burq, N., Gérard, P., Tzvetkov, N., The Cauchy Problem for the nonlinear Schrödinger equation on compact manifolds. In *Phase Space Analysis of Partial Differential Equations* (ed. by F. Colombini and L. Pernazza), vol. I, Centro di Ricerca Matematica Ennio de Giorgi, Scuola Normale Superiore, Pisa 2004, 21–52.

[19] Burq, N., Gérard, P., Tzvetkov, N., Global solutions for the nonlinear Schrödinger equation on three-dimensional compact manifolds. In *Mathematical Aspects of nonlinear PDE*, Annals Math. Studies, Princeton University Press, Princeton, NJ, to appear.

[20] Burq, N., Gérard, P., Tzvetkov, N., An example of singular dynamics for the nonlinear Schrödinger equation on bounded domains. In *Hyperbolic Problems and Related Topics* (ed. by F. Colombini and T. Nishitani), Grad. Ser. Anal., International Press, Somerville, MA, 2003, 57–66.

[21] Burq, N., Planchon, F., Smoothing and dispersive estimates for 1d Schrödinger equations with BV coefficients and applications. Preprint, 2004.

[22] Castro, C., Zuazua, E., Concentration and lack of observability of waves in highly heterogeneous media. *Arch. Ration. Mech. Anal.* **164** (2002), 39–72.

[23] Cazenave, T., *Semilinear Schrödinger equations.* Courant Lect. Notes Math. 10, Amer. Math. Society, Providence, RI, 2003.

[24] Cazenave, T., Weissler, F., The Cauchy problem for the critical nonlinear Schrödinger equation in $H^s$. *Nonlinear Anal.* **14** (1990), 807–836.

[25] Christ, M., Nonuniqueness of weak solutions of the nonlinear Schrödinger equation. Preprint, March 2005.

[26] Christ, M., Colliander, J., Tao, T., Asymptotics, modulation and low regularity ill-posedness for canonical defocusing equations. *Amer. J. Math.* **125** (2003), 1225–1293.

[27] Christ, M., Colliander, J., Tao, T., Ill-posedness for nonlinear Schrödinger and wave equations. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, to appear.

[28] Christ, M., Colliander, J., Tao, T., Instability of the periodic nonlinear Schrödinger equation. Preprint, September 2003.

[29] Colliander, J., Keel, M., Staffilani, G., Takaoka , H., Tao, T., Global wellposedness and scattering in the energy space for the critical nonlinear Schrödinger equation in $\mathbb{R}^3$. *Ann. of Math.*, to appear.

[30] Doi, S. I., Smoothing effects of Schrödinger evolution groups in Riemannian manifolds. *Duke Math. J.* **82** (1996), 679–706.

[31] Gérard, P., Nonlinear Schrödinger equations on compact manifolds. In *European Congress of Mathematics* (Stockholm, 2004), ed. by Ari Laptev, EMS Publishing House, Zürich, 2005, 121–139.

[32] Gérard, P., Pierfelice, V., Nonlinear Schrödinger equation on four-dimensional compact manifolds. Preprint, September 2005.

[33] Ginibre, J., Velo, G., On a class of nonlinear Schrödinger equations. *J. Funct. Anal.* **32** (1979), 1–71.

[34] Ginibre, J., Velo, G., The global Cauchy problem for the nonlinear Schrödinger equation. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **2** (1985) 309–327.

[35] Ginibre, J., Le problème de Cauchy pour des EDP semi-linéaires périodiques en variables d'espace (d'après Bourgain). Séminaire Bourbaki, Exp. 796, *Astérisque* **237** (1996), 163–187.

[36] Hassell, A., Tao, T., Wunsch, J., A Strichartz inequality for the Schrodinger equation on non-trapping asymptotically conic manifolds. *Comm. Partial Differential Equations* **30** (2004), 157–205.

[37] Hassell, A., Tao, T., Wunsch, J., Sharp Strichartz estimates on non-trapping asymptotically conic manifolds. *Amer. J. Math.*, to appear.

[38] Hess, O., Kuhn, T., Maxwell-Bloch equations for spatially inhomogeneous semiconductor lasers. I. Theoretical formulation; II. Spatiotemporal dynamics. *Phys. Rev. A* **54** (1996), 3347–3359; 3360–3368.

[39] Hörmander, L., Estimates for translation invariant operators in $L^p$ spaces. *Acta Math.* **104** (1960), 93–140.

[40] Kappeler, T., Topalov, P., Global wellposedness of the mKdV in $L^2(\mathbb{T}, \mathbb{R})$. *Comm. Partial Differential Equations* **30** (2005), 435–449.

[41] Kato, T., On nonlinear Schrödinger equations. *Ann. Inst. H. Poincaré Phys. Théor.* **46** (1987), 113–129.

[42] Keel, M., Tao, T., Endpoint Strichartz estimates. *Amer. J. Math.* **120** (1998), 955–980.

[43] Kenig, C., Ponce, G., Vega, L., Quadratic forms for $1-D$ semilinear Schrödinger equation. *Trans. Amer. Math. Soc.* **348** (1996), 3323–3353.

[44] Kenig, C., Ponce, G., Vega, L., On the ill-posedness of some canonical dispersive equations. *Duke Math. J.* **106** (2001), 617–633.

[45] Klainerman, S., Machedon, M., Space-time estimates for null forms and the local existence theorem. *Comm. Pure App. Math.* **46** (1993), 1221–1268.

[46] Montgomery-Smith, S.J., Time decay for the bounded mean oscillation of solutions of the Schrödinger and wave equations. *Duke Math. J.* **91** (1998), 393–408.

[47] Ralston, J.V., On the construction of quasimodes associated with stable periodic orbits. *Comm. Math. Phys.* **51** (1976), 219–242; Erratum ibid **67** (1979), 91.

[48] Robbiano, L., Zuily, C., Strichartz estimates for Schrödinger equations with variable coefficients. *Mém. Soc. Math. France* **101–102** (2005).

[49] Salort, D., Dispersion and Strichartz inequalities for the one-dimensional Schrödinger equation with variable coefficients. *Internat. Math. Res. Notices* **11** (2005), 687–700.

[50] Sogge, C., Oscillatory integrals and spherical harmonics. *Duke Math. J.* **53** (1986), 43–65.

[51] Sogge, C., Concerning the $L^p$ norm of spectral clusters for second order elliptic operators on compact manifolds. *J. Funct. Anal.* **77** (1988), 123–138.

[52] Sogge, C., *Fourier integrals in classical analysis.* 105, Cambridge University Press, Cambridge 1993.

[53] Staffilani, G., Tataru, D., Strichartz estimates for a Schrödinger operator with nonsmooth coefficients. *Comm. Partial Differential Equations* **27** (2002), 1337–1372.

[54] Strichartz, R., Restrictions of Fourier transforms to quadratic surfaces and decay of soltions of wave equations. *Duke Math. J.* **44** (1977), 705–714.

[55] Sulem, C., Sulem, P.-L. *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse.* Appl. Math. Sci. 139, Springer-Verlag, New York 1999.

[56] Takaoka, H., Tsutsumi, Y., Well-posedness of the Cauchy problem for the modified KdV equation with periodic boundary conditions. *Internat. Math. Res. Notices* **56** (2004), 3009–3040.

[57] Tataru, D., Strichartz estimates for operators with nonsmooth coefficients and the nonlinear wave equation. *Amer. J. Math.* **122** (2000), 349–376.

[58] Tsutsumi, Y., $L^2$-solutions for nonlinear Schrödinger equations and nonlinear groups. *Funkcial. Ekvac.* **30** (1987), 115–125.

[59] Tzvetkov, N., Illposedness issues for nonlinear dispersive equations. Preprint, September 2004.

[60] Vladimirov, M.V., On the solvability of mixed problem for a nonlinear Schrödinger equation of mixed type. *Soviet Math. Dokl.* **29** (1984), 281–284.

[61] Yajima, K., Existence of solutions for Schrödinger evolution equations, *Comm. Math. Phys.* **110** (1987), 415-426.

[62] Zakharov, V.E., Collapse of Langmuir waves. *Sov. Phys. JETP* **35** (1972), 980–914.

[63] Zygmund, A., On Fourier coefficients of functions of two variables. *Studia Math.* **50** (1974), 189–201.

Laboratoire de Mathématique d'Orsay, UMR 8628 du CNRS, Bâtiment 425,
Université Paris-Sud XI, 91405 Orsay cedex, France

E-mail: patrick.gerard@math.u-psud.fr

# The periodic Lorentz gas in the Boltzmann-Grad limit

François Golse*

**Abstract.** Consider the motion of a single point particle bouncing in a fixed system of spherical obstacles. It is assumed that collisions are perfectly elastic, and that the particle is subject to no external force between collisions, so that the particle moves at constant speed. This type of dynamical system belongs to the class of dispersing billiards, and is referred to as a "Lorentz gas". A Lorentz gas is called periodic when the obstacle centers form a lattice. Assuming that the initial position and direction of the particle are distributed under some smooth density with respect to the uniform measure, one seeks the evolution of that density under the dynamics defined by the particle motion in some large scale limit for which the number of collisions per unit of time is of the order of unity. This scaling limit is known as "the Boltzmann-Grad limit", and is the regime of validity for the Boltzmann equation in the kinetic theory of gases. Whether this evolution is governed in such a limit by a PDE analogous to the Boltzmann equation is a natural question, and the topic of this paper.

**Mathematics Subject Classification (2000).** Primary 82C40, 37A60; Secondary 35B27, 37D50.

**Keywords.** Lorentz gas, dispersing billiards, Boltzmann-Grad limit, kinetic models, mean free path.

## 1. Introduction

In 1905, H. Lorentz proposed the following linear kinetic equation to describe the motion of electrons in a metal [23]:

$$\partial_t f + v \cdot \nabla_x f + \tfrac{1}{m} F(t,x) \cdot \nabla_v f(t,x,v) = N_{\mathrm{at}} r_{\mathrm{at}}^2 |v| \mathcal{C}(f(t,x,\cdot))(v), \qquad (1)$$

where the unknown $f(t,x,v)$ is the density of electrons which, at time $t$, are located at $x$ and have velocity $v$. In (1), $F$ is the electric force field, $m$ the mass of the electron, while $N_{\mathrm{at}}$ and $r_{\mathrm{at}}$ designate respectively the number of metallic atoms per unit volume and the radius of each such atom. Finally $\mathcal{C}(f)$ is the collision integral: it acts on the velocity variable only, and is given, for each continuous $\phi \equiv \phi(v)$ by the formula

$$\mathcal{C}(\phi)(v) = \int_{|\omega|=1, v\cdot\omega>0} \big(\phi(v - 2(v\cdot\omega)\omega) - \phi(v)\big) \cos(v,\omega)\, d\omega. \qquad (2)$$

*The author is grateful to Profs. E. Caglioti, H. S. Dumas and F. Murat for helpful comments during the writing of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Although a microscopic model, this equation is only a statistical description of electron motion and by no means a first principle of electrodynamics. For instance, (1) only holds for probability densities $f$, and does not have distributional solutions of the form

$$f \equiv \delta_{(x(t),v(t))},$$

as one would expect in any situation where there is only one electron and its trajectory in phase space $(x(t), v(t))$ is known exactly (i.e. with probability 1). Obviously, this inconsistency comes from the Lorentz collision integral $\mathcal{C}$, and not from the electric force. Hence we shall assume throughout this lecture that the electric force

$$F \equiv 0$$

and restrict our attention to the collision integral.

Since the Lorentz equation is not itself a first principle of physics, it is natural to understand whether it can be derived from one such first principle. This question belongs to the class of problems known as "hydrodynamic limits" – although in the present case, the term "mesoscopic limit" would be more appropriate.

The interest of mathematicians in this type of question originates in Hilbert's attempts to justify rigorously the equations of fluid mechanics on the basis of the kinetic theory of gases, which he cited as an example in his 6th problem on the axiomatization of physics [19]. In [23], Lorentz himself established his model by analogy with the Boltzmann equation for a gas of hard spheres, and did not seek any rigorous derivation for it – avoiding in particular the rather subtle arguments proposed by L. Boltzmann as a justification for the equation bearing his name.

In this paper, we shall discuss whether the Lorentz equation (1) can be rigorously derived in some asymptotic limit from a very simple mechanistic model for electron motion known as the "Lorentz gas". Although not entirely satisfactory in the context of electrodynamics, this model is to the kinetic theory of electrons what molecular gas dynamics is to the kinetic theory of gases.

## 2. The Lorentz gas

Let[1] $\vec{C} \subset \mathbb{R}^{\mathrm{D}}$ (the dimensions of interest being $\mathrm{D} = 2$ or $\mathrm{D} = 3$) satisfy the condition

$$d(\vec{C}) := \inf_{c,c' \in \vec{C}} |c - c'| > 0. \tag{3}$$

Pick $r \in \left(0, \frac{1}{2} d(\vec{C})\right)$, and consider the motion of a point particle moving at a constant velocity in the domain outside the union of fixed balls of radius $r$ centered at the elements of $\vec{C}$, henceforth denoted

$$Z_r[\vec{C}] := \{x \in \mathbb{R}^{\mathrm{D}} \mid \mathrm{dist}(x, \vec{C}) > r\}. \tag{4}$$

[1] We designate by $\vec{C}$ the set of obstacle centers, to avoid confusion with several constants denoted by $C$ in the sequel.

It is assumed that each collision between the particle and any of the balls is perfectly elastic. Put in other words, denoting by $z$ the collision point and by $n_z$ the inward unit normal to $\partial Z_r[\vec{C}]$ at the point $z$, the pre- and postcollisional velocities $v_-$ and $v_+$ of the particle are related by the Descartes law of specular reflection

$$v_+ = v_- - 2(v_- \cdot n_z)n_z.$$

Obviously, the speed of the particle (i.e. the Euclidian norm of its velocity vector) is invariant under this law of reflection, so that we can assume without loss of generality that this speed is $|v| = 1$.

Assuming that the position and the velocity of the particle are respectively $x$ and $v$ at time $t = 0$, we denote by $X_r(t, x, v; \vec{C})$ and $V_r(t, x, v; \vec{C})$ respectively the position and the velocity of the particle at time $t$. They satisfy the differential equations

$$\begin{aligned} \dot{X}_r &= V_r \quad \text{if } \operatorname{dist}(X(t), \vec{C}) > r, \\ \dot{V}_r &= 0 \quad \text{if } \operatorname{dist}(X(t), \vec{C}) > r, \end{aligned} \tag{5}$$

while

$$\begin{aligned} X_r(t+0) &= X_r(t-0) && \text{if } \operatorname{dist}(X_r(t-0), \vec{C}) = r, \\ V_r(t+0) &= \mathcal{R}[n_{X_r(t-0)}]V_r(t-0) && \text{if } \operatorname{dist}(X_r(t-0), \vec{C}) = r, \end{aligned} \tag{6}$$

where $\mathcal{R}[n]$ designates the specular reflection

$$\mathcal{R}[n]v = v - 2(v \cdot n)n.$$

The dynamical system $(X_r, V_r)$ is referred to as *the Lorentz gas* in the configuration of spherical obstacles of radius $r$ centered at the points of $\vec{C}$.

Let $f^{\text{in}} \equiv f^{\text{in}}(x, v)$ be a probability density on the single-particle phase-space, i.e. a nonnegative measurable function defined a.e. on $Z_r[\vec{C}] \times \mathbb{S}^{D-1}$ such that

$$\iint_{Z_r[\vec{C}]\times\mathbb{S}^{D-1}} f^{\text{in}}(x, v)\, dxdv = 1.$$

Define $f_r \equiv f_r(t, x, v; \vec{C})$ to be the density with respect to $dxdv$ of the image measure of $f^{\text{in}}(x, v)dxdv$ under the flow $(X_r, V_r)$, i.e.

$$f_r(t, x, v; \vec{C}) = f^{\text{in}}(X_r(-t, x, v; \vec{C}), V_r(-t, x, v; \vec{C})). \tag{7}$$

A natural question is whether $f_r(t, x, v; \vec{C})$ converges to a solution of the kinetic equation (1) with $F \equiv 0$ in the vanishing $r$ limit, and under some appropriate scaling assumption on the obstacle configuration $\vec{C}$.

Observe that (5) is the system of ordinary differential equations defining the characteristics of the free transport equation in $Z_r[\vec{C}] \times \mathbb{S}^{D-1}$; therefore the density $f_r$ is the solution of

$$\begin{aligned} \partial_t f_r + v \cdot \nabla_x f_r &= 0, && x \in Z_r[\vec{C}],\ |v| = 1,\ t > 0, \\ f_r(t, x, \mathcal{R}[n_x]v) &= f_r(t, x, v), && x \in \partial Z_r[\vec{C}],\ |v| = 1,\ t > 0, \\ f_r(0, x, v) &= f^{\text{in}}(x, v), && x \in Z_r[\vec{C}],\ |v| = 1. \end{aligned} \tag{8}$$

Hence the question above can be viewed as a some kind of homogenization problem for the transport equation. Analogous homogenization problems for the diffusion (Laplace) equation have been thoroughly studied – the work of Hruslov [20] is one of the first references on this topic; see also the lucid presentation of this class of problems in [10].

G. Gallavotti considered in [14] the case of random configurations $\vec{C}$ of obstacles; specifically, the points in $\vec{C}$ are independent and identically distributed, under Poisson's law with density $N_{\mathrm{at}}$. The radius of the obstacles is $r > 0$; it is assumed that $N_{\mathrm{at}} \to +\infty$ while $r \to 0$ so that $N_{\mathrm{at}} r^2 \to \sigma$. He proved that, in this limit, the expectation of $f_r(t, x, v; \vec{C})$ converges to the solution of (1) with initial data $f^{\mathrm{in}}$ and with $F \equiv 0$. His analysis is written in detail on pp. 48–55 in [15]. Later on, his result was strengthened in [29] by H. Spohn, who considered slightly more general distributions of obstacles. The a.s. convergence of $f_r(t, x, v; \vec{C})$ in $\vec{C}$ was proved by C. Boldrighini, L. A. Bunimovich and Ya. G. Sinai [5].

Obviously, the case of a Poisson distribution of obstacles is very natural in the context of the kinetic theory of (neutral) gases. For instance, one could think of a mixture of two hard sphere gases, one with light molecules, the other one with heavy molecules in equilibrium. If the concentration of the light gas is small, collisions between light molecules can be neglected; only binary collisions involving one molecule of each type are considered. This is essentially[2] the microscopic model studied in [14], [15]. For other applications (such as the motion of electrons in a metal) it may be useful to know what happens for other distributions of obstacles. In this paper, we shall discuss the case of a periodic distribution of obstacles.

## 3. The distribution of free path lengths

From now on, we shall restrict our attention to the case of a periodic Lorentz gas with spherical obstacles of radius $r \in \left(0, \frac{1}{2}\right)$ centered at the integer points, i.e. $\vec{C} = \mathbb{Z}^{\mathrm{D}}$. Since the configuration of obstacle centers is thus fixed, we shall henceforth abbreviate the notation introduced above by setting $X_r(t, x, v) := X_r(t, x, v; \mathbb{Z}^{\mathrm{D}})$, $V_r(t, x, v) := V_r(t, x, v; \mathbb{Z}^{\mathrm{D}})$, while $Z_r := Z_r[\mathbb{Z}^{\mathrm{D}}]$ and $f_r(t, x, v) := f_r(t, x, v, \mathbb{Z}^{\mathrm{D}})$.

In view of the probabilistic interpretation of the kinetic equation (1) and of the definition of the Boltzmann-Grad scaling, one expects that the free path lengths should play an important role in studying the periodic Lorentz gas above in that limit.

**Definition 3.1.** For $x \in Z_r$ and $v \in \mathbb{S}^{\mathrm{D}-1}$, the free path length (or forward exit time) for a particle starting at the position $x$ in the direction $v$ is

$$\tau_r(x, v) = \inf\{t > 0 \mid x + tv \in \partial Z_r\}.$$

[2] Except for the fact that heavy molecules may overlap in Gallavotti's model, while this cannot occur for real hard spheres: see condition (3).

For each $v \in \mathbb{S}^{D-1}$, the function $x \mapsto \tau_r(x, v)$ has a unique continuous extension to $Z_r \cup \{x \in \partial Z_r \mid v \cdot n_x \neq 0\}$ for which we shall abuse the notation $\tau_r(x, v)$.

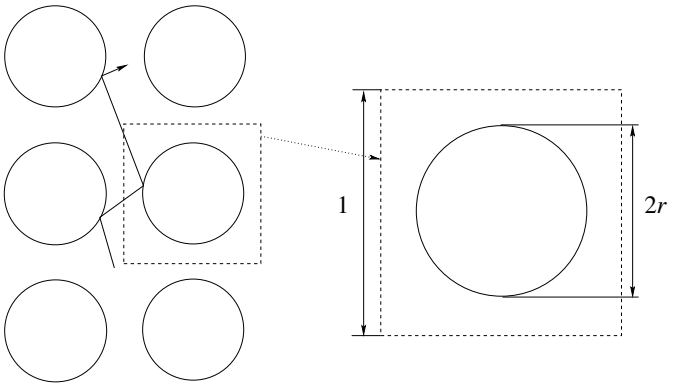



Figure 1. The periodic Lorentz gas.

Notice that $\tau_r(x + k, v) = \tau_r(x, v)$ for each $(x, v) \in Z_r \times \mathbb{S}^{D-1}$ and $k \in \mathbb{Z}^D$: hence $\tau_r$ can be seen as a $[0, +\infty]$-valued function defined on $Y_r \times \mathbb{S}^{D-1}$ and a.e. on $\bar{Y}_r \times \mathbb{S}^{D-1}$, with $Y_r = Z_r/\mathbb{Z}^D$.

If the components of $v \in \mathbb{S}^{D-1}$ are rationally independent – i.e. if $k \cdot v \neq 0$ for each $k \in \mathbb{Z}^D \setminus \{0\}$ – each orbit of the linear flow $x \mapsto x + tv$ is dense on $\mathbb{T}^D = \mathbb{R}^D/\mathbb{Z}^D$, so that $\tau_r(x, v)$ is finite for each $x \in Z_r$.

There are two different, natural phase spaces on which to study the free path length $\tau_r$.

The first one is $\Gamma_r^+ = \{(x, v) \in \partial Z_r \times \mathbb{S}^{D-1} \mid v \cdot n_x > 0\}$ – or its quotient under the action of $\mathbb{Z}^D$-translations on space variables $\tilde{\Gamma}_r^+ = \Gamma_r/\mathbb{Z}^D$ – equipped with its Borel $\sigma$-algebra and the probability measure $\nu_r$ proportional to $\gamma_r$, where

$$d\gamma_r(x, v) = (v \cdot n_x) ds(x) dv,$$

$ds$ being the surface element on $\partial Z_r$.

The second one is $Y_r \times \mathbb{S}^{D-1}$, equipped with its Borel $\sigma$-algebra and the probability measure $\mu_r$ proportional to the uniform measure on $Y_r \times \mathbb{S}^{D-1}$. The measure $\mu_r$ is obviously invariant under the flow ($X$ mod. $\mathbb{Z}^D$, $V$) of the Lorentz gas.

Hence, there are two natural notions of a mean free path for the Lorentz gas:

$$\int_{\tilde{\Gamma}_r^+} \tau_r(x, v)\, d\nu_r(x, v) \quad \text{and} \quad \int_{Y_r \times \mathbb{S}^{D-1}} \tau_r(x, v)\, d\mu_r(x, v). \tag{9}$$

**3.1. Santalò's formula for the mean free path.** In [26], L. Santalò proposed the following simple and elegant explicit expression[3] for the first notion of mean free path.

$$\int_{\tilde{\Gamma}_r^+} \tau_r(x,v)\, d\nu_r(x,v) = \frac{|Y_r||\mathbb{S}^{D-1}|}{\gamma_r(\tilde{\Gamma}_r^+)} = \frac{1-|\mathbb{B}^D|r^D}{|\mathbb{B}^{D-1}|r^{D-1}} \qquad (\textit{Santalò's formula})$$

where $\mathbb{B}^d$ is the d-dimensional unit ball (for the Euclidian norm).

For $D = 3$, one finds

$$\int_{\tilde{\Gamma}_r^+} \tau_r(x,v)\, d\nu_r(x,v) = \frac{1-\frac{4}{3}\pi r^3}{\pi r^2}.$$

With $N_{\mathrm{at}} = 1$ and $|v| = 1$, this is indeed equivalent in the vanishing $r$ limit to the reciprocal of the factor

$$N_{\mathrm{at}} r^2 |v| \int_{|\omega|=1,\, v\cdot\omega>0} \cos(v,\omega)\, d\omega = \pi r^2$$

that appears in (1). However encouraging, this by itself is not enough to justify the relevance of (1) in the description of the Boltzmann-Grad limit of the periodic Lorentz gas (see the discussion in Section 4 below).

Here is a quick proof of Santalò's formula.

**Lemma 3.2** (Dumas–Dumas–Golse [13]). *Let $f \in C^1(\mathbb{R}_+)$ be such that $f(0) = 0$. Then one has*

$$\gamma_r(\tilde{\Gamma}_r^+) \int_{\Gamma_r^+} f(\tau_r(x,v))\, d\nu_r(x,v) = |Y_r||\mathbb{S}^{D-1}| \int_{Y_r\times\mathbb{S}^{D-1}} f'(\tau_r(x,v))\, d\mu_r(x,v)$$

This lemma entails Santalò's formula by letting $f(z) = z$, since the integral on the right-hand side of the identity above is equal to 1.

*Proof.* For each $x \in Z_r$, one has $\tau_r(x+tv,v) = \tau_r(x,v) + t$ for all $t$ near 0. Differentiating in $t$ shows that

$$\begin{aligned} v\cdot\nabla_x\tau_r &= 1, \quad x \in Y_r,\ |v| = 1,\\ \tau_r|_{\tilde{\Gamma}_r^+} &= 0. \end{aligned}$$

Multiplying each side of the first equality by $f'(\tau_r(x,v))$ and integrating for the uniform measure gives

$$\int_{Y_r\times\mathbb{S}^{D-1}} \mathrm{div}_x(vf(\tau_r(x,v)))\, dxdv = |Y_r||\mathbb{S}^{D-1}| \int_{Y_r\times\mathbb{S}^{D-1}} f'(\tau_r(x,v))\, d\mu_r(x,v).$$

We conclude by applying Green's formula to the integral on the left-hand side. □

[3] If $A$ is a d-dimensional measurable subset of $\mathbb{R}^D$ (with $d \le D$), the notation $|A|$ denotes its d-dimensional volume.

**3.2. Bounds on the distribution of free path lengths.** For each point of the form $x = \frac{1}{2}(1, \dots, 1) \in Z_r$, the free path length $\tau_r(x, v)$ is infinite for some $v \in \mathbb{Q}^{\mathrm{D}}$, while it is finite whenever the components of $v$ are not rationally dependent. This suggests the presence of tremendous oscillations in the graph of the function $\tau_r$.

Therefore, it becomes interesting to study the distribution of values of $\tau_r(x, v)$. We shall do so in the phase space $Y_r \times \mathbb{S}^{\mathrm{D}-1}$ equipped with the probability measure $\mu_r$. On the other hand, Santalò's formula suggests that the appropriate scale to measure the free path length is the reciprocal of $r^{\mathrm{D}-1}$. Hence we consider

$$\Phi_r(t) = \mu_r\Big(\Big\{(x, v) \in Y_r \times \mathbb{S}^{\mathrm{D}-1} \;\Big|\; \tau_r(x, v) > \frac{t}{r^{\mathrm{D}-1}}\Big)\Big\}.$$

One could also choose to consider instead

$$\Psi_r(t) = \nu_r\Big(\Big\{(x, v) \in \tilde{\Gamma}_r^+ \;\Big|\; \tau_r(x, v) > \frac{t}{r^{\mathrm{D}-1}}\Big)\Big\}.$$

However, the formula in Lemma 3.2 can be recast in the form

$$\int_0^\infty f(t)\Psi_r(r^{\mathrm{D}-1}t)\,dt = \frac{1 - |\mathbb{B}^{\mathrm{D}}|r^{\mathrm{D}}}{|\mathbb{B}^{\mathrm{D}-1}|r^{\mathrm{D}-1}} \int_0^\infty f'(t)\Phi_r(r^{\mathrm{D}-1}t)\,dt$$

for each $f \in C^1(\mathbb{R}_+)$ such that $f(0) = 0$, which means that

$$\Psi_r = -\tfrac{1-|\mathbb{B}^{\mathrm{D}}|r^{\mathrm{D}}}{|\mathbb{B}^{\mathrm{D}-1}|}\Phi_r' \quad \text{on } \mathbb{R}_+^*. \tag{10}$$

Hence it suffices to study $\Phi_r$.

We begin with the following uniform bounds on $\Phi_r$.

**Theorem 3.3** (Bourgain–Golse–Wennberg [6], [18]). *For any space dimension* D *such that* $\mathrm{D} > 1$*, there exists two positive constants* $C'_{\mathrm{D}} > C_{\mathrm{D}}$ *such that*

$$\frac{C_{\mathrm{D}}}{t} \le \Phi_r(t) \le \frac{C'_{\mathrm{D}}}{t} \quad \textit{for each } t > 1 \textit{ and } r \in \big(0, \tfrac{1}{2}\big).$$

The proof of the upper estimate uses Fourier series in a way that is somewhat reminiscent of Siegel's proof [27] of Minkowski's convex body theorem – see also Theorem 9 in chapter 5 of [24].

The proof of the lower bound is very different in spirit: it is based on a precise counting of infinite open strips included in the billiard table $Z_r$, very similar to Bleher's analysis for the diffusion limit of the periodic Lorentz gas in [2]. Indeed, the free path length $\tau_r(x, v)$ for $x$ in any such strip is bounded from below by the time $\tilde{\tau}_r(x, v)$ at which the trajectory $\{x + tv \mid t > 0\}$ exits the strip. Since $\tilde{\tau}_r(x, v)$ is explicitly known, its distribution is also explicit, and this provides the lower bound for $\Phi_r$.

Since the function $t \mapsto 1/t$ does not belong to $L^1([1, +\infty))$, the lower estimate in Theorem 3.3 implies that the second notion of mean free path in (9) is

$$\int_{Y_r \times \mathbb{S}^{\mathrm{D}-1}} \tau_r(x, v)\,d\mu_r(x, v) = \int_0^\infty \Phi_r(r^{\mathrm{D}-1}t)\,dt = +\infty \quad \text{for each } r \in \big(0, \tfrac{1}{2}\big).$$

**3.3. The distribution of free path lengths for $\mathrm{D} = 2$ as $r \to 0$.** Numerical simulations in [18] suggest that the double inequality in Theorem 3.3 could be strengthened into some asymptotic equivalence as $r \to 0$. However, given the very different nature of the proofs for the upper and the lower bounds in Theorem 3.3, one cannot expect this asymptotic equivalence to be established by the same techniques as in [6].

The proof of Theorem 3.3 suggests that rational approximation plays an important role in the slow decay of the distribution of free path lengths. It is well-known that continued fractions provide a fast algorithm for finding the best rational approximants of any irrational number. For that reason, the Lorentz gas in the case $\mathrm{D} = 2$ can be analyzed in a quite detailed manner with continued fractions, as we shall see below. The same analysis in the case of dimension $\mathrm{D} > 2$ would require using simultaneous rational approximation, a much more difficult problem for which no satisfying analogue of the continued fraction algorithm seems to be available at the time of this writing.

For each $v \in \mathbb{S}^1$, define

$$\phi_r(t, v) = \frac{1}{|Y_r|} \left| \left\{ x \in Y_r \;\middle|\; \tau_r(x, v) > \frac{t}{r^{\mathrm{D}-1}} \right\} \right|.$$

**Theorem 3.4** (Caglioti–Golse [7], [8]). *In the case of space dimension* $\mathrm{D} = 2$*,*

- *for each* $t > 0$ *and a.e.* $v \in \mathbb{S}^1$*,* $\phi_r(t, v)$ *converges in the sense of Cesàro as* $r \to 0$*: there exists* $\phi(t) \in \mathbb{R}_+$ *such that*

$$\lim_{\varepsilon \to 0} \frac{1}{\ln \frac{1}{\varepsilon}} \int_\varepsilon^{r^*} \phi_r(t, v) \, \frac{dr}{r} = \phi(t);$$

- *one has*

$$\phi(t) = \frac{1}{\pi^2 t} + O\Big(\frac{1}{t^2}\Big) \quad \text{as } t \to +\infty.$$

Obviously

$$\Phi_r(t) = \frac{1}{2\pi} \int_{\mathbb{S}^1} \phi_r(t, v) dv, \qquad \text{so that} \qquad \lim_{\varepsilon \to 0} \frac{1}{\ln \frac{1}{\varepsilon}} \int_\varepsilon^{r^*} \Phi_r(t) \, \frac{dr}{r} = \phi(t). \tag{11}$$

The asymptotic expansion $\frac{1}{\pi^2 t} + O\big(\frac{1}{t^2}\big)$ has been identified for the first time in [7]. In fact, the result in [7] stated that both the lim sup and the lim inf of the Cesàro mean of $\Phi_r$ for the scaling invariant measure $\frac{dr}{r}$ as in (11) have that same asymptotic expansion. The a.e. pointwise (in $v$) convergence is new – see [8].

**3.3.1. Method of proof.** Before sketching the proof of the result above, let us recall some background on continued fractions.

The Gauss map is defined as

$$T : (0, 1) \setminus \mathbb{Q} \ni x \mapsto \frac{1}{x} - \left[ \frac{1}{x} \right] \in (0, 1) \setminus \mathbb{Q};$$

it is an ergodic automorphism of $(0,1)\setminus\mathbb{Q}$ with respect to the Gauss measure $dg(x) = \frac{1}{\ln 2}\frac{dx}{1+x}$ that is invariant under $T$.

Let $x \in (0,1)\setminus\mathbb{Q}$; define the sequence of positive integers

$$a_k = \left[\frac{1}{T^{k-1}x}\right], \quad k \geq 1.$$

Then $x$ is represented by the continued fraction

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} =: [0; a_1, a_2, a_3, \dots].$$

Define by induction the sequences of integers $p_n$ and $q_n$ by

$$\begin{aligned} p_{n+1} &= a_n p_n + p_{n-1}, \quad \text{for each } n \geq 1, \quad p_0 = 1,\ p_1 = 0, \\ q_{n+1} &= a_n q_n + q_{n-1}, \quad \text{for each } n \geq 1, \quad q_0 = 0,\ q_1 = 1, \end{aligned} \tag{12}$$

For each $n \geq 2$, the integers $p_n$ and $q_n$ are coprime, and the rational number $\frac{p_n}{q_n}$ is called the $n$-th convergent of $x$. The distance from $x$ to its $n$-th convergent is measured by

$$d_n = (-1)^{n-1}(q_n x - p_n) > 0; \tag{13}$$

for each $n \geq 0$, one has

$$d_n = \prod_{k=0}^{n-1} T^k x. \tag{14}$$

(see for instance the third formula on p. 89 of [28]).

*Step* 1. *A three-term partition of* $\mathbb{T}^2$. A key idea in the proof of Theorem 3.4 is provided by the answer found by S. Blank and N. Krikorian [1] to the following question raised by R. Thom: *"What is the longest orbit of a linear flow with irrational slope on a flat torus with a disk removed?"*

Without loss of generality, assume that the linear flow is $x \mapsto x + tv$ with $v = (\cos\theta, \sin\theta)$ and $\theta \in (0, \frac{\pi}{4})$. The removed disk of radius $r$ is then replaced with the vertical slit $S_r(v)$ of length $2r/\cos\theta$ as shown in Figure 2 (left). Blank and Krikorian found that the set of lengths of all orbits of the linear flow above on $\mathbb{T}^2 \setminus S_r(v)$ consists of exactly three positive values, $l_A(r,v) < l_B(r,v)$ and $l_C(r,v) = l_A(r,v) + l_B(r,v)$.

This suggests considering the three-term partition of $\mathbb{T}^2 \setminus S_r(v)$

$$\{Y_A(r,v), Y_B(r,v), Y_C(r,v)\}$$

defined as follows: $Y_A(r,v)$ (resp. $Y_B(r,v)$, $Y_C(r,v)$) is the union of all orbits of length $l_A(r,v)$ (resp. $l_B(r,v)$, $l_C(r,v)$). Set

$$S_A(r,v) = \{y \in S_r(v) \mid \text{ the orbit starting from } y \text{ is of type } A\}$$

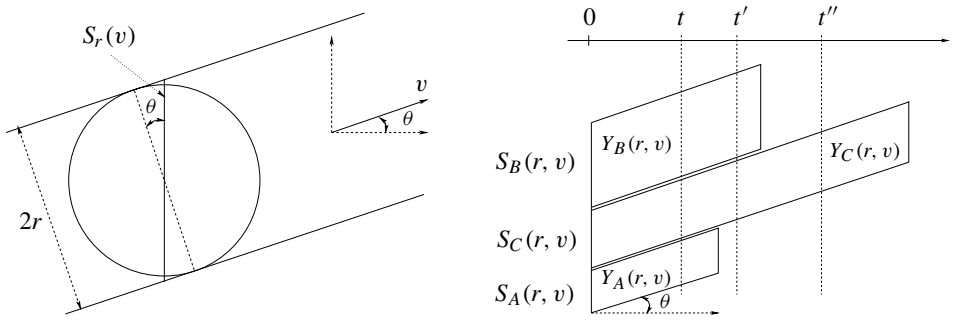



Figure 2. Left: The obstacle and the slit. Right: The 3-strip partition of the 2-torus. This figure gives a simple geometric interpretation of $\psi_r(rs, v)$ for $rs\cos\theta = t$, $rs\cos\theta = t'$ or $rs\cos\theta = t''$.

with analogous definitions for $S_B(r,v)$ and $S_C(r,v)$. Then $S_A(r,v)$, $S_B(r,v)$ and $S_C(r,v)$ are segments while $Y_A(r,v)$, $Y_B(r,v)$ and $Y_C(r,v)$ are (mod 1) parallelograms with one side being $S_A(r,v)$, $S_B(r,v)$ or $S_C(r,v)$ while the adjacent sides are of lengths $l_A(r,v)$, $l_B(r,v)$, or $l_C(r,v)$: see Figure 2 (right).

The orbit lengths $l_A(r,v)$, $l_B(r,v)$ and $l_C(r,v)$, and the lengths of the three segments $S_A(r,v)$, $S_B(r,v)$ and $S_C(r,v)$ are computed in terms of $r$ and the continued fraction expansion of $\tan\theta$ as follows.

Set $\alpha = \tan\theta$, and denote by $\alpha = [0; a_1(\theta), a_2(\theta), a_3(\theta), \dots]$ the continued fraction expansion of $\alpha = \tan\theta$, also let $p_n(\alpha)/q_n(\alpha)$ be the $n$-th convergent of $\alpha$ as in (12). Finally, let $d_n(\alpha)$ be the sequence of errors as defined in (13).

Define

$$N(\alpha, r) = \min\left\{n \in \mathbb{N} \mid d_n(\alpha) \le 2r\sqrt{1+\alpha^2}\right\}; \tag{15}$$

and

$$k(\alpha, r) = -\left[\frac{2r\sqrt{1+\alpha^2} - d_{N(\alpha,r)-1}}{d_{N(\alpha,r)}}\right] \tag{16}$$

Then, the three-strip partition above is characterized by the formulas below:

$$\begin{aligned} l_A(r,v) &= q_{N(\alpha,r)}(\alpha)\sqrt{1+\alpha^2}, \\ l_B(r,v) &= \left(q_{N(\alpha,r)-1}(\alpha) + k(\alpha,r)q_{N(\alpha,r)}(\alpha)\right)\sqrt{1+\alpha^2}, \\ l_C(r,v) &= \left(q_{N(\alpha,r)-1}(\alpha) + (k(\alpha,r)+1)q_{N(\alpha,r)}(\alpha)\right)\sqrt{1+\alpha^2}, \end{aligned} \tag{17}$$

while

$$\begin{aligned} |S_A(\alpha,r)| &= 2r\sqrt{1+\alpha^2} - d_{N(\alpha,r)}(\alpha), \\ |S_B(\alpha,r)| &= 2r\sqrt{1+\alpha^2} - \left(d_{N(\alpha,r)-1}(\alpha) - k(\alpha,r)d_{N(\alpha,r)}(\alpha)\right), \\ |S_C(\alpha,r)| &= d_{N(\alpha,r)-1}(\alpha) - (k(\alpha,r)-1)d_{N(\alpha,r)}(\alpha) - 2r\sqrt{1+\alpha^2}. \end{aligned} \tag{18}$$

*Step* 2. *Computing* $\phi_r$. Let $\lambda_r(x, v) = \inf\{t > 0 \mid x + tv \in S_r(v)\}$ for each $x \in \mathbb{T}^2 \setminus S_r(v)$; clearly

$$|\tau_r(x, v) - \lambda_r(x, v)| \leq r \quad \text{for each } x \in Y_r \setminus S_r(v). \tag{19}$$

Define

$$\psi_r(t, v) := \text{Prob}\{x \in \mathbb{T}^2 \setminus S_r(v) \mid \lambda_r(x, v) \geq t/r\},$$

where the probability is computed with respect to the uniform measure on $Y_r$. Because of (19), one has

$$\psi_r(t - r^2, v) - \pi r^2 \leq (1 - \pi r^2)\phi_r(t, v) \leq \psi_r(t + r^2, v) \quad \text{for each } t \geq r^2. \tag{20}$$

On the other hand, $\psi_r$ can be computed explicitly with the help of the three-term partition above. It is found that

$$\begin{aligned}\psi_r(t, v) = \max\big(1 - 2t,\, 1 - \tfrac{1-\delta_N}{\delta_{N-1}}\mu_N - 2t\delta_N,\, 1 - \tfrac{(k-1)\delta_N+1}{\delta_{N-1}}\mu_N - \tfrac{\delta_N}{\delta_{N-2}}\mu_{N-1} \\ - (\delta_{N-1} - (k-1)\delta_N - 1)\big(2t - \tfrac{\mu_{N-1}}{\delta_{N-2}} - k\tfrac{\mu_N}{\delta_{N-1}}\big),\, 0\big).\end{aligned} \tag{21}$$

In the formula above, $N = N(\alpha, r)$ and $\delta_n = \frac{d_n(\alpha)}{2r\sqrt{1+\alpha^2}}$ while $\mu_n = d_{n-1}(\alpha)q_n(\alpha)$; also $k = k(\alpha, r) = -\big[-\big(\frac{\delta_{N-1}}{\delta_N} - \frac{1}{\delta_N}\big)\big]$. The direction is $v = \big(\frac{1}{\sqrt{1+\alpha^2}}, \frac{\alpha}{\sqrt{1+\alpha^2}}\big)$.

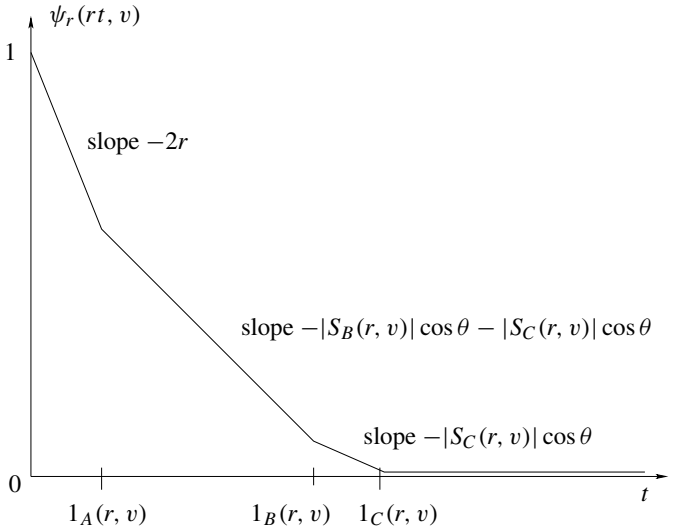



Figure 3. Graph of $\psi_r(rt, v)$.

*Step* 3. *Using the ergodicity of the Gauss map.* Birkhoff's ergodic theorem says that, for each $h \in L^1\big((0,1), \frac{dx}{1+x}\big)$, one has

$$\frac{1}{N}\sum_{j=0}^{N-1} h(T^m\alpha) \to \frac{1}{\ln 2}\int_0^1 \frac{h(z)dz}{1+z} \quad \text{a.e. in } \alpha \in (0,1) \text{ as } N \to +\infty. \tag{22}$$

Together with formula (14) and the definition of $N(\alpha, r)$, the convergence in (22) for $h = \ln$ implies that

$$N(\alpha, r) \sim \tfrac{12\ln 2}{\pi^2}|\ln r| \quad \text{as } r \to 0, \text{ for a.e. } \alpha \in (0,1). \tag{23}$$

Define

$$\Delta_j(\alpha, x) := -\ln \delta_{N(\alpha,e^{-x})-j}(\alpha) = -\ln d_{N(\alpha,e^{-x})-j}(\alpha) - x + \ln(2\sqrt{1+\alpha^2}) \tag{24}$$

for each $j \geq 0$, $\alpha \in (0,1) \setminus \mathbb{Q}$ and $x > \ln 2$. A further application of Birkhoff's theorem (22) leads to

**Lemma 3.5.** *Let $f$ be a bounded continuous function on $\mathbb{R}^{m+1}$. Then, for each $x_* \geq \ln 2$, one has*

$$\frac{1}{\ln(1/r)}\int_{x^*}^{\ln(1/r)} f(\Delta_0(\alpha,x), \dots, \Delta_m(\alpha,x))dx \to \int_0^1 \frac{F(\theta)d\theta}{1+\theta} \quad \textit{a.e. in } \alpha \in (0,1)$$

*as $r \to 0$, where*

$$F(\theta) = \int_0^{|\ln(T^m\theta)|} f(Y_m(y,\theta))\, dy.$$

*In the formula above, $Y_m(y,\theta)$ denotes*

$$Y_m(y,\theta) = (y,\, y + \ln T^m\theta,\, y + \ln T^m\theta + \ln T^{m-1}\theta,\, \dots,\, y + \ln T^m\theta + \cdots + \ln T\theta).$$

*Step* 4. *The small scatterer limit for the Cesàro mean of $\Phi_r$.* We seek to apply the lemma above to compute

$$\frac{1}{\ln(1/\varepsilon)}\int_\varepsilon^{1/2} \psi_r(t,v)\,\frac{dr}{r} \quad \text{in the limit as } \varepsilon \to 0.$$

Unfortunately, $\psi_r$ given by (21) is not a function of any fixed, finite number of ratios of the form $\frac{\delta_n}{\delta_{n-1}}$, but also involves a few $\mu_n$s – in the original variables, $\psi_r$ explicitly depends on the $q_n(\alpha)$s which involve the complete string of all the $T^j\alpha$s for $j = 0, 1, \dots, n-1$, not only the last one.

Next observe that $\frac{\mu_N}{\delta_{N-1}} \leq 1$, and hence $t \geq 1$ implies that

$$\begin{aligned}\psi_r(t,v) &= \max\big(1 - \tfrac{1-\delta_N}{\delta_{N-1}}\mu_N - 2t\delta_N,\, 1 - \tfrac{(k-1)\delta_N+1}{\delta_{N-1}}\mu_N - \tfrac{\delta_N}{\delta_{N-2}}\mu_{N-1} \\ &\qquad - (\delta_{N-1} - (k-1)\delta_N - 1)\big(2t - \tfrac{\mu_{N-1}}{\delta_{N-2}} - k\tfrac{\mu_N}{\delta_{N-1}}\big),\, 0\big) \\ &= \max\big(1 - \tfrac{1-\delta_N}{\delta_{N-1}}\mu_N - 2t\delta_N, \\ &\qquad (\delta_{N-1} - (k-1)\delta_N - 1)\big(\tfrac{\mu_{N-1}}{\delta_{N-2}} + (k+1)\tfrac{\mu_N}{\delta_{N-1}} - 2t\big),\, 0\big).\end{aligned} \tag{25}$$

In this last equality we have used formula (8) in chapter 1 of [21].

On the other hand, $\delta_{N-1} - (k-1)\delta_N - 1 \leq \delta_N$ so that

$$\begin{aligned} 0 \leq \psi_r(t,v) - (1 - \tfrac{1-\delta_N}{\delta_{N-1}}\mu_N - 2t\delta_N)_+ \\ \leq \tfrac{\delta_N}{\delta_{N-1}}\mu_N \mathbf{1}_{2t \leq \frac{\mu_{N-1}}{\delta_{N-2}} + (k+1)\frac{\mu_N}{\delta_{N-1}}} \leq \tfrac{\delta_N}{\delta_{N-1}}\mu_N \mathbf{1}_{2t \leq (k+2)\frac{\mu_N}{\delta_{N-1}}} \leq \tfrac{1}{k}\mathbf{1}_{k+2 \geq t}. \end{aligned}$$

Finally $\frac{\delta_N}{\delta_{N-1}}\mu_N \leq \frac{1}{k}$ and $1 - \mu_N \leq \frac{2}{k}$ (see Lemma 4.1 in [7]) so that

$$\begin{aligned} \left|\left(1 - \tfrac{1-\delta_N}{\delta_{N-1}}\mu_N - 2t\delta_N\right)_+ - \left(1 - \tfrac{1}{\delta_{N-1}} - 2t\delta_N\right)_+\right| &\leq \left(\tfrac{\delta_N}{\delta_{N-1}}\mu_N + \tfrac{1-\mu_N}{\delta_{N-1}}\right)\mathbf{1}_{k+2 \geq t} \\ &\leq \tfrac{3}{k}\mathbf{1}_{k+2 \geq t} \end{aligned}$$

and $\psi_r(t,v)$ can be replaced with $(1 - \frac{1}{\delta_{N-1}} - 2t\delta_N)_+$ modulo an error term controlled by $\frac{3}{k}\mathbf{1}_{k+2 \geq t}$. Applying Lemma 3.5 to $f(\Delta_0, \Delta_1) = (1 - e^{\Delta_1} - 2te^{-\Delta_0})_+$ leads to the asymptotic behavior in the second part of Theorem 3.4.

The proof of the a.e. in $v$ convergence uses Steps 1–3 above, in a way that is somehow more involved: see [8] for more details.

**3.3.2. Later improvements.** Theorem 3.4 was later strengthened by F. Boca and A. Zaharescu [4], in two different ways. First, they were able to remove the need for Cesàro averaging in the convergence statement of (11). Also, they obtained a (semi-)explicit formula for $\phi$. Here is their result:

**Theorem 3.6** (Boca–Zaharescu [4]). *In the case of space dimension* $\mathrm{D} = 2$*, one has*

$$\Phi_r(t) \to \phi(t) \quad \text{as } r \to 0 \text{ for each } t > 0$$

*where*

$$\begin{aligned} \phi(t) &= 1 - 2t + \tfrac{12}{\pi^2}t^2, && t \in \left(0, \tfrac{1}{2}\right], \\ \phi(t) &= \tfrac{6}{\pi^2}\int_0^{2t-1} a(x,t)\,dx + \tfrac{6}{\pi^2}\int_{2t-1}^1 b(x,t)\,dx, && t \in \left(\tfrac{1}{2}, 1\right], \\ \phi(t) &= \tfrac{6}{\pi^2}\int_0^1 a(x,t)\,dx, && t \in (1, +\infty), \end{aligned}$$

*with the functions a and b given by*

$$\begin{aligned} a(x,t) &= \tfrac{(1-x)^2}{x}\left(2\ln\tfrac{2t-x}{2(t-x)} - \tfrac{2t}{x}\ln\tfrac{(2t-x)^2}{4t(t-x)}\right), \\ b(x,t) &= \tfrac{1-2t}{x}\ln\tfrac{1}{2t-x} + \tfrac{(2t-x)(x+1-2t)}{x} + \tfrac{(1-x)^2}{x}\left(2\ln\tfrac{2t-x}{1-x} - \tfrac{2t}{x}\ln\tfrac{(2t-x)}{2t(1-x)}\right). \end{aligned}$$

The formulas above for $\Phi$ were first conjectured by P. Dahlqvist in [11], by an argument involving Farey fractions, which however remained incomplete since it ultimately relied on the equidistribution of a certain geometrical quantity, which remained to be proved.

The proof by Boca and Zaharescu is essentially based on two ideas: a) using the same 3-strip partition as in [7], in the language of Farey instead of continuous fractions, and b) computing certain sums indexed by lattice points with coprime coordinates by replacing them with integrals while controlling the resulting error terms.

However, being based on averaging in $x$ and $v$, their proof fails to provide a.e. pointwise convergence in $v$, unlike the proof of Theorem 3.4, based on Birkhoff's ergodic theorem for the Gauss map, which requires instead averaging in $r$, thereby proving only convergence in Cesàro's sense.

**3.4. The entropy of the billiard map as $r \to 0$.** The semi-explicit formula for $\Phi$ in Theorem 3.6 has at least one important application besides the problem of justifying the Lorentz equation (1). Define the *billiard map* in the case of the Lorentz gas to be

$$\mathcal{B}_r : \tilde{\Gamma}_r^+ \to \tilde{\Gamma}_r^+, \quad (x, v) \mapsto \mathcal{B}_r(x, v) = (x + \tau_r(x, v)v, \mathcal{R}[n_{x+\tau_r(x,v)v}]v); \quad (26)$$

one easily checks that the measure $\nu_r$ is invariant under the map $\mathcal{B}_r$. Denote by $h(\mathcal{B}_r)$ the Kolmogorov–Sinai entropy of the billiard map $\mathcal{B}_r$ with respect to the measure $\nu_r$. A consequence of Theorem 3.6 and of formula (10)[4] is the following asymptotic formula for the entropy of the billiard map in dimension $\mathrm{D} = 2$ and in the small obstacle limit:

$$h(\mathcal{B}_r) = 2 \ln \frac{1}{r} + 2 + C + o(1) \quad \text{as } r \to 0.$$

Here the constant $C$ is defined as

$$C = \lim_{r \to 0} \left( \int_{\tilde{\Gamma}_r^+} \ln \tau_r(x, v)\, d\nu_r(x, v) - \ln \int_{\tilde{\Gamma}_r^+} \tau_r(x, v)\, d\nu_r(x, v) \right) = \frac{9\zeta(3)}{4\zeta(2)} - 3 \ln 2$$

while $\zeta$ is Riemann's zeta function.

In 1991, N. Chernov had proved that, in dimension D, the entropy of the billiard map satisfies

$$h(\mathcal{B}_r) = \mathrm{D}(\mathrm{D} - 1) \ln \frac{1}{r} + O(1) \quad \text{as } r \to 0;$$

see [9] and the references therein. That the $O(1)$ error term should actually converge as $r \to 0$ had been conjectured earlier by B. Friedman, Y. Oono and I. Kubo on the basis of numerical simulations; the correct value of the limit was then proposed by Dahlqvist in [11] before Boca–Zaharescu's proof in [4].

## 4. The Boltzmann-Grad limit: a negative result

In this section, we return to the formulation of the Boltzmann-Grad limit for the periodic Lorentz gas in terms of a homogenization problem for the transport equation, as in Section 2.

[4] In [4], Boca and Zaharescu do not use formula (10); instead they derive the formula for the distribution $\Psi_r$ by using again the 3-term partition and the approximation of sums over coprime lattice points as in the proof of Theorem 3.6.

Going back to the free transport equation (8), we set

$$\vec{C} := \varepsilon\mathbb{Z}^{\mathrm{D}}, \quad r_\varepsilon = \varepsilon^{\frac{\mathrm{D}}{\mathrm{D}-1}}, \quad \Omega_\varepsilon := Z_{r_\varepsilon}[\varepsilon\mathbb{Z}^{\mathrm{D}}] \quad \text{and} \quad f_\varepsilon(t,x,v) := f_{r_\varepsilon}(t,x,v;\varepsilon\mathbb{Z}^{\mathrm{D}}),$$

where $\varepsilon \in (0, 2^{-\mathrm{D}})$. Hence $f_\varepsilon$ satisfies

$$\begin{aligned} &\partial_t f_\varepsilon + v\cdot\nabla_x f_\varepsilon = 0, && x\in\Omega_\varepsilon,\ v\in\mathbb{S}^{\mathrm{D}-1},\\ &f_\varepsilon(t,x,v) = f_\varepsilon(t,x,\mathcal{R}[n_x]v), && x\in\partial\Omega_\varepsilon,\ v\in\mathbb{S}^{\mathrm{D}-1}. \end{aligned} \tag{27}$$

For simplicity, we shall assume that $\varepsilon$ is of the form $\varepsilon = \frac{1}{n}$ for $n > 2^{\mathrm{D}}$, and that the initial data $f^{\mathrm{in}} \equiv f^{\mathrm{in}}(x,v)$ is continuous and periodic in $x$ with period 1 in each coordinate direction. In other words, $f^{\mathrm{in}} \in C(\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1})$.

With the choice of $\varepsilon = \frac{1}{n}$, the solution $f_\varepsilon$ of (27) with initial data

$$f_\varepsilon(0,x,v) = f^{\mathrm{in}}(x,v), \quad x\in\Omega_\varepsilon,\ v\in\mathbb{S}^{\mathrm{D}-1}, \tag{28}$$

is also periodic in the variable $x$ with period 1 in each coordinate direction; if one extends $f_\varepsilon$ by 0 inside the obstacles and abuse the notation $f_\varepsilon$ to designate this extension, one sees that

$$f_\varepsilon \in L^\infty(\mathbb{R}_+\times\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1}) \quad \text{with } \|f_\varepsilon\|_{L^\infty} = \|f^{\mathrm{in}}\|_{L^\infty}.$$

By the Banach–Alaoglu theorem, the sequence $f_\varepsilon$ (for $\varepsilon = \frac{1}{n}$ with $n > 2^{\mathrm{D}}$) is relatively compact in $L^\infty(\mathbb{R}_+\times\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1})$ for the weak-* topology. It is therefore natural to investigate the limit points of $f_\varepsilon$ as $\varepsilon\to 0$ – this being exactly the Boltzmann-Grad limit of the periodic Lorentz gas viewed as a homogenization problem for the transport equation.

We begin with the following negative result:

**Theorem 4.1** (Golse [17]). *There exists initial data $f^{\mathrm{in}} \in L^\infty(\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1})$ such that no subsequence of $f_\varepsilon$ converges in $L^\infty(\mathbb{R}_+\times\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1})$ weak-* to the solution of the Lorentz kinetic equation* (1).

In fact, the result in [17] is stronger: it excludes the possibility that any subsequence of $f_\varepsilon$ converges in $L^\infty(\mathbb{R}_+\times\mathbb{T}^{\mathrm{D}}\times\mathbb{S}^{\mathrm{D}-1})$ weak-* to the solution of any linear Boltzmann equation of the form

$$(\partial_t + v\cdot\nabla_x) f(t,x,v) = \sigma\int_{\mathbb{S}^{\mathrm{D}-1}} k(v,v')(f(t,x,v') - f(t,x,v))\,dv' \tag{29}$$

with $\sigma > 0$ and $k \in C(\mathbb{S}^{\mathrm{D}-1}\times\mathbb{S}^{\mathrm{D}-1})$ such that

$$k(v,w) = k(w,v) > 0, \quad \int_{\mathbb{S}^{\mathrm{D}-1}} k(v,w)\,dv = 1.$$

The proof is based on the fact that the operator

$$f \mapsto v\cdot\nabla_x f + \sigma\int_{\mathbb{S}^{\mathrm{D}-1}} k(v,v')(f(t,x,v) - f(t,x,v'))\,dv'$$

with domain

$$\{f \equiv f(x,v) \in L^2(\mathbb{T}^D \times \mathbb{S}^{D-1}) \mid v \cdot \nabla_x f \in L^2(\mathbb{T}^D \times \mathbb{S}^{D-1})\}$$

is Fredholm with nullspace the set of constant functions. Hence there exists $c > 0$ such that

$$\begin{aligned} &\left\| f(t,\cdot,\cdot) - \tfrac{1}{|\mathbb{S}^{D-1}|} \iint_{\mathbb{T}^D \times \mathbb{S}^{D-1}} f(t,y,w)\, dy dw \right\|_{L^2(\mathbb{T}^D \times \mathbb{S}^{D-1})} \\ &\qquad \le C \| f|_{t=0} \|_{L^2(\mathbb{T}^D \times \mathbb{S}^{D-1})} e^{-ct} \end{aligned} \tag{30}$$

for each solution of (29). On the other hand, if $f^{\mathrm{in}} \ge 0$ a.e., the solution $f_\varepsilon$ of (27) satisfies

$$f_\varepsilon(t,x,v) \ge f^{\mathrm{in}}(x - tv, v) \mathbf{1}_{t \le \varepsilon \tau_\varepsilon^{1/(D-1)}(x,v)}$$

– the right-hand side being the solution of the same transport equation as in (27) but with absorbing boundary condition

$$f_\varepsilon(t,x,v) = 0 \quad \text{for } x \in \partial\Omega_\varepsilon \text{ and } v \cdot n_x > 0.$$

Hence, if $f$ is any weak-* limit point of $f_\varepsilon$ in $L^\infty(\mathbb{R}_+ \times \mathbb{T}^D \times \mathbb{S}^{D-1})$ as $\varepsilon \to 0$, it must satisfy

$$f(t,x,v) \ge \frac{C_D}{t} f^{\mathrm{in}}(x - tv, v)$$

by Theorem 3.3. This is incompatible with (30) as can be seen by taking $f^{\mathrm{in}}(x,v) \equiv \rho(x)$ with $\|\rho\|_{L^2(\mathbb{T}^D)} = 1$ while $\|\rho\|_{L^1(\mathbb{T}^D)} = o(1)$.

The case of a Lorentz gas with purely absorbing obstacles is much simpler and yet not without interest. Let $g_\varepsilon \equiv g_\varepsilon(t,x,v)$ be the solution of

$$\begin{aligned} \partial_t g_\varepsilon + v \cdot \nabla_x g_\varepsilon &= 0, && x \in \Omega_\varepsilon,\ v \in \mathbb{S}^{D-1},\ t > 0, \\ g_\varepsilon(t,x,v) &= 0, && x \in \partial\Omega_\varepsilon,\ v \cdot n_x > 0, \\ g_\varepsilon(0,x,v) &= f^{\mathrm{in}}(x,v), && x \in \Omega_\varepsilon,\ v \in \mathbb{S}^{D-1}. \end{aligned} \tag{31}$$

In the 2-dimensional case, Theorem 3.4 provides a complete description of the limit:

**Theorem 4.2** (Caglioti–Golse [7], [8]). *For each $f^{\mathrm{in}} \in L^\infty(\mathbb{T}^2 \times \mathbb{S}^1)$,*

$$\frac{1}{\ln \frac{1}{\eta}} \int_\eta^{1/2} g_\varepsilon \frac{d\varepsilon}{\varepsilon} \to g$$

*weakly-* in $L^\infty(\mathbb{T}^2 \times \mathbb{S}^1)$ and pointwise in $t \ge 0$ as $\varepsilon \to 0$, with $g$ given by*

$$g(t,x,v) = f^{\mathrm{in}}(x - tv, v)\phi(t),$$

*where $\phi$ is the small scatterer limit of the distribution of free path lengths, whose explicit expression is provided by Theorem* 3.6.

In other words, $g$ is the solution of

$$\begin{aligned} \partial_t g + v \cdot \nabla_x g &= \tfrac{\phi'(t)}{\phi(t)} g, \quad x \in \mathbb{R}^2,\ v \in \mathbb{S}^1,\ t > 0, \\ g_\varepsilon|_{t=0} &= f^{\mathrm{in}}. \end{aligned} \tag{32}$$

Notice that $\frac{\phi'(t)}{\phi(t)} < 0$, so that the term on the right-hand side of (32) indeed models the loss of particles impinging on the obstacles.

This result can be viewed as a homogenization problem for the free transport equation in a domain with holes. The analogous problem for the diffusion (Laplace) equation has been analyzed in detail: see for instance [20], [10]. It describes the steady, D-dimensional motion of particles on Brownian trajectories in a periodic array of circular holes with radius $\varepsilon^{\frac{\mathrm{D}}{\mathrm{D}-2}}$ centered at the points of the cubic lattice $\varepsilon\mathbb{Z}^{\mathrm{D}}$, each particle falling into a hole being permanently removed. Notice the different critical size of the obstacles – $\varepsilon^{\frac{\mathrm{D}}{\mathrm{D}-2}}$ in the diffusion case, instead of $\varepsilon^{\frac{\mathrm{D}}{\mathrm{D}-1}}$ in the transport case – which comes from the fact that the diffusion and free transport operators are of order 2 and 1 respectively, thereby leading to different scalings. More importantly, in the case of the diffusion problem, the loss of particles falling into the holes is described in this limit with a constant absorption coefficient. Indeed, successive increments in Brownian trajectories are independent random variables, so that the periodic structure of the array of holes is somehow ignored by the particles. On the contrary, in the case of the free transport problem (31), the trajectories are straight lines, which introduces correlations between the obstacles. Intuitively, particles which have not encountered any obstacle over some interval of time $[0, T]$ move in a direction that is well approximated by a rational direction – with increasing quality of approximation as $T$ increases. Such particles are therefore much less likely to encounter obstacles after time $T$, and this agrees with the fact that the absorption rate $\frac{\phi'(t)}{\phi(t)}$ vanishes as $t \to +\infty$.

## 5. Conclusion

The methods presented above explain why the Lorentz kinetic equation (1) fails to describe the Lorentz gas in the Boltzmann-Grad limit, when the obstacles are centered at the vertices of the cubic lattice $\mathbb{Z}^{\mathrm{D}}$. The ergodic theory of continued fractions provides additional insight on this example of periodic Lorentz gas in the case $\mathrm{D} = 2$, especially on the asymptotic distribution of free path lengths in the small obstacle limit.

Obviously, it would be desirable to obtain as much information in higher dimensions, particularly for the physically relevant case $\mathrm{D} = 3$. This could be difficult, as it might require accurate estimates on simultaneous rational approximation.

Otherwise, it would be useful to have analogues of the results above for 2-dimensional lattices other than $\mathbb{Z}^2$. Specifically, one would like to know whether

Theorems 3.3, 3.4 and 3.6 can be extended or adapted to the case of arbitrary 2-dimensional lattices. If so, it would be particularly interesting to find the intrinsic meaning of the constants $\frac{1}{\pi^2}$ and $\frac{9\zeta(3)}{4\zeta(2)} - 3\ln 2$ that appear in Theorem 3.4 and in Section 3.4.

Finally, the problem of finding an equation describing the Boltzmann-Grad limit of the periodic Lorentz gas – even in the simplest 2-dimensional case and for the cubic lattice $\mathbb{Z}^2$ – remains open. So far, we have no clue as to the structure of such an equation, should it exist: we only know that it cannot be a linear Boltzmann equation of the type (29).

## References


[1] Blank, S. J., Krikorian, N., Thom's problem on irrational flows. *Internat. J. Math.* **4** (1993), 721–726.

[2] Bleher, P., Statistical properties of two-dimensional periodic Lorentz gas with infinite horizon. *J. Statist. Phys.* **66** (1992), 315–373.

[3] Bobylev, A. V., Hansen, A., Piasecki, J., Hauge, E. H., From the Liouville equation to the generalized Boltzmann equation for magneto-transport in the 2D Lorentz model. *J. Statist. Phys.* **102** (2001), 1133–1150.

[4] Boca, F., Zaharescu, A., The distribution of the free path lengths in the periodic two-dimensional Lorentz gas in the small scatterer limit. Preprint.

[5] Boldrighini, C., Bunimovich, L. A., Sinai, Ya. G., On the Boltzmann equation for the Lorentz gas. *J. Statist. Phys.* **32** (1983), 477–501.

[6] Bourgain, J., Golse, F., Wennberg, B., On the distribution of free path lengths in the periodic Lorentz gas. *Comm. Math. Phys.* **190** (1998), 491–508.

[7] Caglioti, E., Golse, F., On the Distribution of Free Path Lengths in the Periodic Lorentz Gas III. *Comm. Math. Phys.* **236** (2003), 199–221.

[8] Caglioti, E., Golse, F., The Boltzmann-Grad Limit of the Billiard Map for the 2D Periodic Lorentz Gas. In preparation.

[9] Chernov, N. I., Entropy values and entropy bounds. In *Hard Ball Systems and the Lorentz Gas* (ed. by D. Szász), Encyclopaedia Math. Sci. 101, Springer-Verlag, Berlin 2000, 121–143.

[10] Cioranescu, D., Murat, F. Un terme étrange venu d'ailleurs. In *Nonlinear Partial Differential Equations and their Applications*, Collège de France Seminar, Vol. II (Paris, 1979/1980), Res. Notes in Math. 60, Pitman, Boston, Mass./London 1982, 98–138, 389–390.

[11] Dahlqvist, P., The Lyapunov exponent in the Sinai billiard in the small scatterer limit. *Nonlinearity* **10** (1997), 159–173.

[12] Desvillettes, L., Ricci, V., Nonmarkovianity of the Boltzmann-Grad limit of a system of random obstacles in a given force field. *Bull. Sci. Math.* **128** (2004), 39–46.

[13] Dumas, H. S., Dumas, L., Golse, F., Remarks on the notion of mean free path for a periodic array of spherical obstacles. *J. Statist. Phys.* **87** (1997), 943–950.

[14] Gallavotti, G., Divergences and approach to equilibrium in the Lorentz and the wind-tree models. *Phys. Rev.* (2) **185** (1969), 308–322.

[15] Gallavotti, G., *Statistical Mechanics. A Short Treatise*. Texts and Monographs in Physics, Springer-Verlag, Berlin, Heidelberg 1999.

[16] Golse, F., Hydrodynamic Limits. In *European Congress of Mathematics* (Stockholm, 2004), ed. by A. Laptev, European Math. Soc. Publishing House, Zürich 2005, 669–717.

[17] Golse, F., On the periodic Lorentz gas in the Boltzmann-Grad scaling. Preprint.

[18] Golse, F., Wennberg, B., On the Distribution of Free Path Lengths in the Periodic Lorentz Gas II. *M2AN Modél. Math. et Anal. Numér.* **34** (2000), 1151–1163.

[19] Hilbert, D., Mathematical Problems. International Congress of Mathematicians (Paris, 1900); translated and reprinted in *Bull. Amer. Math. Soc.* **37** (2000), 407–436.

[20] Hruslov, Ē. Ja., The method of orthogonal projections and the Dirichlet boundary value problem in domains with a "fine-grained" boundary. *Mat. Sb.* (*N.S.*) **88** (130) (1972), 38–60; English transl. *Math. USSR Sb.* **17** (1972), 37–59.

[21] Khinchin, A. Ya., *Continued Fractions*. The University of Chicago Press, Chicago, Ill., London, 1964

[22] Lanford, Oscar E., III., Time evolution of large classical systems. In *Dynamical Systems, Theory and Applications* (Rencontres, Battelle Res. Inst., Seattle, Wash., 1974), Lecture Notes in Phys. 38, Springer-Verlag, Berlin 1975, 1–111.

[23] Lorentz, H., Le mouvement des électrons dans les métaux. *Arch. Néerl.* **10** (1905), 336–371.

[24] Montgomery, Hugh L. *Ten Lectures on the Interface between Analytic Number Theory and Harmonic Analysis*. CBMS Reg. Conf. Ser. Math. 84, Amer. Math. Soc., Providence, RI, 1994.

[25] Ricci, V., Wennberg, B., On the derivation of a linear Boltzmann equation from a periodic lattice gas. *Stochastic Process. Appl.* **111** (2004), 281–315.

[26] Santalò, L. A., Sobre la distribución probable de corpúsculos en un cuerpo, deducida de la distribución en sus secciones y problemas analogos. *Revista Union Mat. Argentina* **9** (1943), 145–164.

[27] Siegel, C. L., Über Gitterpunkte in convexen Körpern und ein damit zusammenhängendes Extremalproblem. *Acta Math.* **65** (1935), 307–323.

[28] Sinaĭ, Ya. G., *Topics in Ergodic Theory*. Princeton Math. Ser. 44, Princeton University Press, Princeton, NJ, 1994.

[29] Spohn, H., The Lorentz process converges to a random flight process. *Comm. Math. Phys.* **60** (1978), 277–290 .

Université Paris 7-Denis Diderot, Laboratoire J.-L. Lions, Boîte courrier 187, 4 place Jussieu, 75252 Paris Cedex 05, France

E-mail: golse@math.jussieu.fr

# Conformal invariants and nonlinear elliptic equations

Matthew J. Gursky

**Abstract.** We describe several "uniformizing" problems in conformal geometry, all of which can be formulated as problems of existence for solutions of certain elliptic partial differential equations. For the sake of exposition we divide the discussion according to the type of PDE, beginning with semilinear equations related to the scalar curvature, then higher order equations arising from the $Q$-curvature, and finally fully nonlinear equations.

**Mathematics Subject Classification (2000).** Primary 58J05; Secondary 53A30.

**Keywords.** Conformal geometry, Yamabe equation, Paneitz operator, $\sigma_k$-curvature.

## 1. Introduction

In this article we outline several problems related to finding a canonical representative of a conformal equivalence class of Riemannian metrics. The earliest and best known result of this kind is the Uniformization Theorem for surfaces. Although originally conceived as a problem in complex function theory, in view of the modern development of the theory of Riemann surfaces it can be stated in purely geometric terms: Given a compact surface $(M^2, g)$, there is a conformal metric $\tilde{g} = e^{2w}g$ with constant curvature.

In higher dimensions the Uniformization Theorem is no longer true, even locally. For example, in dimensions $n \geq 4$ the Weyl tensor $W(g)$ provides an obstruction to a metric being locally conformal to a flat metric. Despite this fact – or perhaps because of it – there are numerous ways to define what constitutes a canonical metric. In this article we will outline several "uniformizing" problems, each involving the study of a PDE of a different type: second order semilinear (in the case of the scalar curvature), higher order semilinear (for the $Q$-curvature), and fully nonlinear (when considering the $\sigma_k$-curvature). An underlying theme will be the connection between conformal invariants associated to the problem, spectral properties of the relevant differential operator, and the interplay of both with the topology of the manifold.

## 2. Semilinear examples

**2.1. The modified scalar curvature.** Let $(M^n, g)$ be a closed Riemannian manifold of dimension $n \geq 3$, and let $R(g)$ denote its scalar curvature. If $\tilde{g} = u^{4/(n-2)}g$ is a

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

conformal metric, then the scalar curvature $R(\tilde{g})$ of $\tilde{g}$ is given by

$$L_g u + R(\tilde{g}) u^{\frac{n+2}{n-2}} = 0, \tag{2.1}$$

where $L_g = \frac{4(n-1)}{(n-2)}\Delta_g - R(g)$ is the *conformal laplacian*. The *Yamabe problem* is to establish the existence of a conformal metric with constant scalar curvature (see [32]). Since the scalar curvature of a surface is twice the Gauss curvature, the Yamabe problem can be viewed as one possible generalization of the Uniformization Theorem.

An important property of the operator $L_g$ is its *conformal covariance*: If $\tilde{g} = u^{4/(n-2)} g$, then

$$L_{\tilde{g}} v = u^{-\frac{n+2}{n-2}} L_g(uv). \tag{2.2}$$

Consequently, the sign of the principle eigenvalue $\lambda_1(-L_g)$ is a conformal invariant ([30]).

A generalization of the Yamabe problem was introduced by the author, for the purposes of studying a rigidity question for Einstein metrics [22]. Subsequently it has been used to estimate the Yamabe invariant of certain four-manifolds [23], the principal eigenvalue of the Dirac operator [28], and in the context of Seiberg–Witten theory [31].

Given a Riemannian manifold $(M^n, g)$, let $\mathcal{G} \subset S^2T^*M^n$ denote the ray bundle consisting of metrics in the conformal class of $g$. Let $\delta_s : \mathcal{G} \to \mathcal{G}$ denote the dilations $\delta_s(g) = s^2 g$, with $s > 0$. Functions on $\mathcal{G}$ which are homogeneous of degree $\beta$ with respect to $\delta_s$ are known as *conformal densities of weight* $\beta$.

**Definition 2.1.** Given a density $\phi$ of weight $-2$, we define the *modified scalar curvature* associated to $\phi$ by

$$\hat{R}(g) = R(g) - \phi. \tag{2.3}$$

For example, in some applications $(M^4, g)$ is an oriented four-dimensional manifold and $\phi = 2\sqrt{6}|W^{\pm}(g)|$, the (anti-) self-dual part of the Weyl curvature tensor. For Kähler manifolds of positive scalar curvature the corresponding modified scalar curvature $\hat{R}(g) = R(g) - 2\sqrt{6}|W^+|^2 \equiv 0$. As this example illustrates, in general we do not assume the density is smooth, but at least Lipschitz continuous.

Since $\phi$ is a conformal density of weight $-2$ the operator $\hat{L}_g = \frac{4(n-1)}{(n-2)}\Delta_g - \hat{R}(g)$ enjoys the same invariance as the conformal laplacian given in (2.2). In particular, the sign of $\lambda_1(-\hat{L}_g)$ is a conformal invariant.

In analogy with the Yamabe problem, we can ask whether there exists a conformal metric whose modified scalar curvature is constant. Due to the conformal covariance of $\hat{L}$, the sign of this constant will agree with the sign of $\lambda_1(-\hat{L}_g)$. Since the choice of density obviously plays an important role in the question of existence, a completely general theory would seem unlikely. For particular cases, though, some existence results have been appeared (see [29]).

In fact, for many applications the relevant question is the *sign* of the modified scalar curvature, or equivalently, the sign of $\lambda_1(-\hat{L}_g)$. Typically, the density $\phi$ is chosen by examining the curvature term in the Weitzenböck formula for a harmonic section of some vector bundle; then the sign of $\lambda_1(-\hat{L}_g)$ can be thought of as an obstruction to the existence of non-trivial harmonic sections. Of course, Lichnerowicz used precisely this kind of argument with the Dirac operator on a spin manifold to prove obstructions to the existence of metrics with positive scalar curvature [33].

To illustrate this technique with one important example, consider a self-dual harmonic two-form $\omega \in H^2_+(M^4, \mathbb{R})$. In this case the Weitzenböck formula is given by

$$\frac{1}{2}\Delta_g|\omega|^2 = |\nabla\omega|^2 - 2W^+(\omega,\omega) + \frac{1}{3}R(g)|\omega|^2. \tag{2.4}$$

Another intriguing element in the study of the modified scalar curvature is the important role played by *refined Kato inequalities* (see [24]). In the case of self-dual harmonic two-forms this takes the form

$$|\nabla\omega|^2 \geq \frac{3}{2}|\nabla|\omega||^2 \tag{2.5}$$

([37]). Substituting into (2.4), and using the fact that $W^+\colon \Lambda^2_+ \to \Lambda^2_+$ is a trace-free endomorphism, we eventually arrive at

$$\Delta_g|\omega|^{2/3} \geq \frac{1}{6}\big[-2\sqrt{6}|W^+(g)| + R(g)\big]|\omega|^{2/3}. \tag{2.6}$$

Taking $\phi = 2\sqrt{6}|W^+|$, we conclude

$$\hat{L}_g|\omega|^{2/3} \geq 0 \tag{2.7}$$

which implies $\lambda_1(-\hat{L}_g) \leq 0$. Thus, there are cohomological obstructions to the existence of metrics with positive first eigenvalue.

On the other hand, when $\lambda_1(-\hat{L}_g) \leq 0$ one obtains $L^p$-estimates for the scalar curvature. For example, in four dimensions

$$\int R(g)^2\, dv(g) \leq \int \phi^2\, dv(g). \tag{2.8}$$

Since $\phi$ is a density of weight $-2$, the integral on the right-hand side of (2.8) is a conformal invariant. Thus, we have a connection between the spectral properties of the conformally covariant operator $\hat{L}_g$, and $L^2$-conformal invariants. We will encounter a similar phenomenon when considering higher order elliptic equations.

**2.2. Higher order equations and the $Q$-curvature.** In an unpublished preprint [35], the late Stephen Paneitz constructed a fourth order conformally covariant operator defined on a (pseudo)-Riemannian manifold $(M^4, g)$ of dimension $n \geq 3$. In

four dimensions, his operator is given by

$$P_g = (-\Delta_g)^2 - \delta_g\left\{\left[\frac{2}{3}R(g)g - 2\,\mathrm{Ric}(g)\right]\circ\nabla\right\}, \tag{2.9}$$

where $\delta_g\colon \Lambda^1(M^4) \to C^\infty(M^4)$ is the divergence. If $\tilde{g} = e^{2w}g$ is a conformal metric, then

$$P_{\tilde{g}} = e^{-4w}P_g. \tag{2.10}$$

In particular, the sign of $\lambda_1(P_g)$ and the kernel of $P_g$ are both conformally invariant. Since $P_g(1) = 0$, we always have $\lambda_1(P_g) \leq 0$.

Branson [4] subsequently observed the connection between Paneitz's operator and what Branson called the *Q-curvature*, defined by

$$Q(g) = \frac{1}{12}\big(-\Delta_g R(g) + R(g)^2 - 3|\,\mathrm{Ric}(g)|^2\big). \tag{2.11}$$

In fact, if $\tilde{g} = e^{2w}g$, then the $Q$-curvature of $\tilde{g}$ is given by

$$P_g w + 2Q(g) = 2Q(\tilde{g})e^{4w}. \tag{2.12}$$

It follows that the integral of the $Q$-curvature is another conformal invariant:

$$\int Q(\tilde{g})\,dv(\tilde{g}) = \int Q(g)\,dv(g). \tag{2.13}$$

The $Q$-curvature and Paneitz operator have become important objects of study in the geometry of four-manifolds, and play a role in the such diverse topics as the Moser–Trudinger inequalities ([3], [5]), twistor theory ([14]), gauge choices for Maxwell's equations ([13]), and conformally compact AHE manifolds ([15], [16]). In addition, they naturally suggest another "uniformizing" problem, that of finding a conformal metric with constant $Q$-curvature.

Chang and Yang ([9]) studied this problem using the direct variational method, by attempting to prove the existence of minimizers of the non-convex functional

$$\begin{aligned} F[w] = &\int w P_g w\,dv(g) + 4\int wQ(g)\,dv(g) \\ &- \left(\int Q(g)\,dv(g)\right)\log\int e^{4w}\,dv(g). \end{aligned} \tag{2.14}$$

However, if the Paneitz operator has a negative eigenvalue and the conformal invariant (2.13) is positive, then Chang and Yang showed that $\inf F = -\infty$. Even when $F$ is bounded below the compactness of a minimizing sequence is a delicate matter. Using a sharp form of Adam's inequality ([1]) Chang and Yang [9] were able to prove the existence of minimizer assuming the invariant (2.13) is less than its value on the standard sphere:

**Theorem 2.2.** *Let* $(M^4, g)$ *be a closed four-manifold, and assume* (i) $P_g \geq 0$ *with* $\operatorname{Ker} P_g = \{\text{const.}\}$, *and* (ii) $\int Q(g)dv(g) < 8\pi^2$. *Then there exists a minimizer* $w \in C^\infty$ *of* $F$, *which satisfies* (2.12) *with* $Q(\tilde{g}) = \text{const.}$

Thus, the question of existence is reduced to studying the spectrum of the Paneitz operator and the conformal invariant (2.13). As we shall see, there is a connection between these problems.

First, it is not difficult to see that when $g$ has positive scalar curvature, the invariant (2.13) is *always* dominated by its value on the sphere. (Somewhat surprisingly, equality can be characterized without resorting to the Positive Mass Theorem; see [21]). Thus, when $(M^4, g)$ has positive scalar curvature the assumption $(ii)$ in the Theorem of Chang-Yang is superfluous, except in the case of the sphere.

Turning to the first assumption of Theorem 2.2, we begin by noting the Dirichlet form associated to the Paneitz operator is given by

$$\int \psi P_g \psi \, dv(g) = \int \left[ (\Delta_g \psi)^2 + \frac{2}{3} R(g) |\nabla \psi|^2 - 2 \operatorname{Ric}(g)(\nabla \psi, \nabla \psi) \right] dv(g). \tag{2.15}$$

Using the Bochner formula we can rewrite this as

$$\begin{aligned} \int \psi P_g \psi \, dv(g) &= \int \left[ \frac{4}{3} |\mathring{\nabla}^2 \psi|^2 \, dv(g) + \frac{2}{3} \big( R(g) g - \operatorname{Ric}(g) \big)(\nabla \psi, \nabla \psi) \right] dv(g) \\ &\geq \int \frac{4}{3} |\mathring{\nabla}^2 \psi|^2 \, dv(g) + \int \frac{2}{3} T(\nabla \psi, \nabla \psi) \, dv(g), \end{aligned} \tag{2.16}$$

where $\mathring{\nabla}^2 \psi = \nabla^2 \psi - \frac{1}{4}(\Delta_g \psi) g$ is the trace-free Hessian and $T = R(g)g - \operatorname{Ric}(g)$. Consequently, if the right-hand side of (2.16) is positive for all $\psi \in C^\infty$, then $\lambda_1(P_g) = 0$ and $\operatorname{Ker} P_g = \{\text{const.}\}$. By conformal invariance it is enough to show that this property holds for some metric in the conformal class of $g$.

By using a kind of "regularized" version of the functional $F$ (which was also studied by Chang and Yang), the author was able to construct a metric $\tilde{g} \in [g]$ for which the right-hand side of (2.16) is positive for all $\psi \in C^\infty$, provided the scalar curvature of $g$ is positive (see [21]):

**Theorem 2.3.** *If* $(M^4, g)$ *has positive scalar curvature and* $\int Q(g) \, dv(g) \geq 0$, *then* $P_g \geq 0$ *and* $\operatorname{Ker} P_g = \{\text{const.}\}$. *In particular, there is a conformal metric* $\tilde{g}$ *with* $Q(\tilde{g}) = \text{const.}$

The assumptions of Theorem 2.3 imply the first Betti number of $M^4$ vanishes; see [20]. On the other hand, in [14] Eastwood–Singer constructed metrics on $k(S^3 \times S^1)$ for all $k > 0$ with $P_g \geq 0$ and $\operatorname{Ker} P_g = \{\text{const.}\}$. For this reason, it would be desirable to relax the assumption on the integral of $Q(g)$. A result of this kind appears in [25], which relied on solving a fully nonlinear equation of the type described in the next section to prove the positivity of the tensor field $T = R(g)g - \operatorname{Ric}(g)$ appearing in (2.16):

**Theorem 2.4.** *If $(M^4, g)$ has positive Yamabe invariant $Y(g)$ and the $Q$-curvature satisfies $\int Q(g)\, dv(g) + \frac{1}{6}Y(g)^2 \geq 0$, then $P_g \geq 0$ and $\operatorname{Ker} P_g = \{\text{const.}\}$. In particular, there is a conformal metric $\tilde{g}$ with $Q(\tilde{g}) = \text{const.}$*

Since Theorem 2.4 allows the integral of $Q$ to be negative, we are able to construct many new examples of conformal manifolds which admit a metric with constant $Q$-curvature (see Section 7 of [25].

There have been other approaches to the problem of finding metrics with constant $Q$-curvature. Malchiodi [34], Malchiodi–Djadli [11], and Druet–Robert [12] have studied existence and compactness of the solution space by a delicate blow-up analysis. In particular, the positivity of the total $Q$-curvature is not required. However, their work indicates that the assumption $\operatorname{Ker} P_g = \{\text{const.}\}$ is unlikely to be merely technical. Brendle [7] and Baird–Fardoun–Regbauoi [2] have used parabolic methods; they also assume $P_g \geq 0$ with trivial kernel.

The problem of finding metrics with constant $Q$-curvature originally appeared in the more general context of studying variational properties of the regularized determinant (see [9]). In contrast to the study of the $Q$-curvature, the existence theory has not developed very much beyond the results in [9] and [21]. The associated Euler–Lagrange equation includes terms which are nonlinear in the second derivatives of the solution, providing an important link to the material in the next section.

## 3. Fully nonlinear equations

Finally, we give a brief synopsis of a very active area which can be viewed as a fully nonlinear version of the Yamabe problem. For Riemannian manifolds of dimension $n \geq 3$ we define the *Weyl–Schouten* tensor by

$$A(g) = \frac{1}{(n-2)}\left(\operatorname{Ric}(g) - \frac{1}{2(n-1)}R(g)g\right). \tag{3.1}$$

In [39], J. Viaclovsky initiated the study of the fully nonlinear equations arising from the transformation of $A$ under conformal deformations. More precisely, let $g_u = e^{-2u}g$ denote a conformal metric, and consider the equation

$$\sigma_k^{1/k}(g_u^{-1} \circ A_u) = f(x), \tag{3.2}$$

where $\sigma_k(\cdot)$ denotes the $k$-th elementary symmetric polynomial, applied to the eigenvalues of $g_u^{-1} \circ A_u$. Since $A_u$ is related to $A$ by the formula

$$A(g_u) = A(g) + \nabla^2 u + du \otimes du - \frac{1}{2}|\nabla u|^2 g, \tag{3.3}$$

equation (3.2) is equivalent to

$$\sigma_k^{1/k}(A(g) + \nabla^2 u + du \otimes du - \frac{1}{2}|\nabla u|^2 g) = f(x)e^{-2u}. \tag{3.4}$$

Note that when $k = 1$, $\sigma_1(g^{-1}A(g)) = \text{trace}(A) = \frac{1}{2(n-1)}R(g)$; therefore, (3.4) is equivalent to equation (2.1). When $k > 1$ the equation is fully nonlinear, but not necessarily elliptic. A sufficient condition for a solution $u \in C^2(M^n)$ to be elliptic is that the eigenvalues of $A = A(g)$ are in $\Gamma_k^+ = \{\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n \mid \sigma_1(\lambda) > 0, \sigma_2(\lambda) > 0, \ldots, \sigma_k(\lambda) > 0\}$ at each point of $M^n$. In this case we say that $g$ is *admissible* (or $k$-admissible); likewise, if $-A(g) \in \Gamma_k^+$ then we say $g$ is *negative admissible*.

The most straightforward question one can pose about equation (3.4) is existence: given an admissible metric $g \in \Gamma_k^+(M^n)$ and a positive function $f \in C^\infty(M^n)$, does there always exist a solution $u \in C^\infty(M^n)$ to (3.4)? When $f(x) = \text{const.} > 0$ this is referred to as the *$\sigma_k$-Yamabe problem*. An important distinction between the $\sigma_k$-Yamabe problem and the classical Yamabe problem is that the former is not in general variational.

The study of equation (3.4) in general and the $\sigma_k$-Yamabe problem in particular has seen an explosion of activity in recent years. We will highlight some results for the case $k > n/2$ as the theory is more developed; admissibility has a more geometric interpretation; and in contrast to the case $k \leq n/2$, solutions of (3.4) have a variational characterization.

First, Guan–Viaclovsky–Wang [19] showed that when $g$ is $k$-admissible the Ricci curvature satisfies the sharp inequality

$$\text{Ric}(g) \geq \frac{(2k-n)}{2n(k-1)}R(g)g. \tag{3.5}$$

In particular, if $k > n/2$, the Ricci curvature is positive. Using the Newton–Maclaurin inequality and Bishop's volume comparison, one can quantify this in the following way: If $g_u = e^{-2u}g$ is an admissible solution of (3.4) with $f(x) \geq c_0 > 0$, then

$$\text{Vol}(g_u) = \int e^{-nu}\, dv(g) \leq C(k, n, c_0). \tag{3.6}$$

We define the *$k^{\text{th}}$-maximal volume* of the admissible metric $g$ by

$$\Lambda_k(M^n, [g]) = \sup\{\text{Vol}(g_u) \mid A(g_u) \in \Gamma_k^+, \sigma_k(g_u^{-1}A(g_u)) \geq \sigma_k(S^n)\}, \tag{3.7}$$

where $\sigma_k(S^n)$ is the value attained by the round sphere. This definition suggests the following variational problem: given a $k$-admissible metric $g$, is there is conformal metric which attains the $k^{\text{th}}$-maximal volume? It is easy to see that any metric which does will satisfy (3.4) with $f(x) \equiv \sigma_k^{1/k}(S^n)$. A similar variational problem was formulated by Guan and Spruck in studying the curvature of hypersurfaces in Euclidean space [17].

In joint work with J. Viaclovsky we used this variational scheme to study the $\sigma_k$-Yamabe problem in three- and four-dimensions [26]. The dimension restriction is a result of the difficulty of proving sharp estimates for the maximal volume in high dimensions. In three dimensions we could prove such estimates thanks to Bray's

Football Theorem [6], and in four dimensions by using the Chern–Gauss–Bonnet formula. An approach somewhat similar in spirit was implicit in earlier work of Chang–Gursky–Yang [8] and Viaclovsky [40].

Another consequence of the volume bound (3.6) is the finiteness of the blow-up set for a sequence of solutions to (3.4). This follows from the $\varepsilon$-regularity result of Guan and Wang [18], which in turn is based on their local $C^1$- and $C^2$-estimates for solutions. In fact, using the estimates of Guan-Wang, it is possible to show that a divergent sequence $\{u_i\}$ of solutions to (3.4) will blow up at finitely many points, and converge uniformly to $-\infty$ off the singular set. By rescaling this sequence, one obtains a limiting viscosity solution $w \in C^{1,1}_{\text{loc}}$ with $f(x) \geq 0$.

In recent work with Viaclovsky [27], we carried out a careful analysis of the tangent cone at infinity of the $C^{1,1}$-metric $g_\infty = e^{2w}g$. In particular, we showed the volume growth at infinity is Euclidean. Since the Ricci curvature is non-negative, this implies the metric is flat, and $(M^n, g)$ is conformally the sphere.

**Theorem 3.1.** *Let $(M^n, g)$ be closed $n$-dimensional Riemannian manifold, and assume*

(i) *$g$ is $k$-admissible with $k > n/2$, and*

(ii) *$(M^n, g)$ is not conformally equivalent to the round $n$-dimensional sphere.*

*Then given any smooth positive function $f \in C^\infty(M^n)$ there exists a solution $u \in C^\infty(M^n)$ of* (3.4)*, and the set of all such solutions is compact in the $C^m$-topology for any $m \geq 0$.*

In fact, our proof gives the existence of solutions to

$$F(A_u) = f(x)e^{-2u},$$

where $F\colon \Gamma \to \mathbb{R}$ is a symmetric function of the eigenvalues of $A_u$ defined on a cone $\Gamma \subset \mathbb{R}^n$ which satisfies some explicit structural conditions. For more general equations we need to use the $C^2$-estimates developed by S. Chen [10]. Trudinger and Wang [38] have proved a similar existence result, along with a remarkable Harnack inequality for admissible metrics.

## References

[1] Adams, D. R., A sharp inequality of J. Moser for higher order derivatives. *Ann. of Math.* **128** (2) (1988), 385–398.

[2] Baird, P., Fardoun, A., Regbauoi, Q-curvature flow on 4-manifolds. *Calc. Var. Partial Differential Equations*, to appear.

[3] Beckner, W., Sharp Sobolev inequalities on the sphere and the Moser-Trudinger inequality. *Ann. of Math.* **138** (1) (1993), 213–242.

[4] Branson, T. P., Differential operators canonically associated to a conformal structure. *Math. Scand.* **57** (1985), 293–345.

[5] Branson, T. P., Chang, S.-Y. A., Yang, P. C., Estimates and extremals for zeta function determinants on four-manifolds. *Comm. Math. Phys.* **149** (2) (1992), 241–262.

[6] Bray, H., The Penrose inequality in General Relativity and volume comparison theorems involving scalar curvature. Dissertation, Stanford University, 1997.

[7] Brendle, S., Global existence and convergence for a higher order flow in conformal geometry. *Ann. of Math.* **158** (1) (2003), 323–343.

[8] Chang, S.-Y. A., Gursky, J., Yang, C., An a priori estimate for a fully nonlinear equation on four-manifolds. *J. Anal. Math.* **87** (2002), 151–186.

[9] Chang, S.-Y. A., Yang, P. C., Extremal metrics of zeta function determinants on 4-manifolds. *Ann. of Math.* **142** (1995), 171–212.

[10] Chen, S., Local estimates for some fully nonlinear elliptic equations. *Internat. Math. Res. Notices* **2005** (2005), 3403–3425.

[11] Djadli, Z., Malchiodi, A., A fourth order uniformization theorem on some four manifolds with large total $Q$-curvature. *C. R. Math. Acad. Sci. Paris* **340** (5) (2005), 341–346.

[12] Druet, O., Robert, F., Bubbling phenomena for fourth-order four-dimensional PDEs with exponential growth. *Proc. Amer. Math. Soc.* **134** (2006), 897–908.

[13] Eastwood, M., Singer, M., A conformally invariant Maxwell gauge. *Phys. Lett. A* **107** (2) (1985), 73–74.

[14] Eastwood, M. G., Singer, M. A., The Fröhlicher spectral sequence on a twistor space. *J. Differential Geom.* **38** (1993), 653–669.

[15] Fefferman, C., Graham, C. R., $Q$-curvature and Poincaré metrics. *Math. Res. Lett.* **9** (2–3) (2002), 139–151.

[16] Graham, C. R., Zworski, M., Scattering matrix in conformal geometry. *Invent. Math.* **152** (1) (2003), 89–118.

[17] Guan, B., Spruck, J., Locally convex hypersurfaces of constant curvature with boundary. *Comm. Pure Appl. Math.* **57** (10) (2004), 1311–1331.

[18] Guan, Pengfei, Wang, Guofang, Local estimates for a class of fully nonlinear equations arising from conformal geometry. *Internat. Math. Res. Notices* **2003** (2003), 1413–1432.

[19] Guan, P., Viaclovsky, J. A., Wang, G., Some properties of the Schouten tensor and applications to conformal geometry. *Trans. Amer. Math. Soc.* **355** (2003), 925–933.

[20] Gursky, M. J., The Weyl functional, de Rham cohomology, and Kähler-Einstein metrics. *Ann. of Math.* **148** (1) (1998), 315–337.

[21] Gursky, M. J., The principal eigenvalue of a conformally invariant differential operator, with an application to semilinear elliptic PDE. *Comm. Math. Phys.* **207** (1999), 131–143.

[22] Gursky, M. J., Four-manifolds with $\delta W^+ = 0$ and Einstein constants of the sphere. *Math. Ann.* **318** (2000), 417–431.

[23] Gursky, M. J., LeBrun, C., Yamabe invariants and spin$^c$ structures. *Geom. Funct. Anal.* **8** (1998), 965–977.

[24] Gursky, M. J., LeBrun, C., On Einstein manifolds of positive sectional curvature. *Ann. Global Anal. Geom.* **17** (1999), 315–328.

[25] Gursky, M. J., and Viaclovsky, J. A., A fully nonlinear equation on four-manifolds with positive scalar curvature. *J. Differential Geom.* **63** (2003), 131–154.

[26] Gursky, M. J., Viaclovsky, J. A., Volume comparison and the $\sigma_k$-Yamabe problem. *Adv. Math.* **187** (2004), 447–487.

[27] Gursky, M. J., Viaclovsky, J. A., Prescribing symmetric functions of the eigenvalues of the Ricci tensor. *Ann. of Math.*, to appear.

[28] Herzlich, M., Moroianu, A., Generalized Killing spinors and conformal eigenvalue estimates for Spin$^c$ manifolds. *Ann. Global Anal. Geom.* **17** (4) (1999), 341–370.

[29] Itoh, M., The modified Yamabe problem and geometry of modified scalar curvatures. *J. Geom. Anal.* **15** (1) (2005), 63–81.

[30] Kazdan, J. L., Warner, F. W., Scalar curvature and conformal deformation of Riemannian structure. *J. Differential Geom.* **10** (1975), 113–134.

[31] LeBrun, C., Ricci curvature, minimal volumes, and Seiberg-Witten theory. *Invent. Math.* **145** (2) (2001), 279–316.

[32] Lee, J. M., Parker, T. H., The Yamabe problem, *Bull. Amer. Math. Soc.* (*N.S.*) **17** (1987), 37–91.

[33] Lichnerowicz, A., Spineurs harmoniques. *C. R. Acad. Sci. Paris* **239** (1963), 7–9.

[34] Malchiodi, A., Compactness of solutions to some geometric fourth-order equations. *J. Reine Angew. Math.*, to appear.

[35] Paneitz, S., A quartic conformally covariant differential operator for arbitrary pseudo-Riemannian manifolds. Preprint, 1983.

[36] Schoen, Richard M., Variational theory for the total scalar curvature functional for Riemannian metrics and related topics. In *Topics in Calculus of Variations* (Montecatini Terme, 1987), Lecture Notes in Math. 1365, Springer-Verlag, Berlin 1989, 120–154.

[37] Seaman, W., Harmonic two-forms in four dimensions. *Proc. Amer. Math. Soc.* **112** (2) (1991), 545–548.

[38] Trudinger, N. S., Wang, X.-J., On Harnack inequalities and singularities of admissible metrics in the Yamabe problem. Preprint, 2005.

[39] Viaclovsky, J. A., Conformal geometry, contact geometry, and the calculus of variations. *Duke Math. J.* **101** (2000), 283–316.

[40] Viaclovsky, J. A., Estimates and existence results for some fully nonlinear elliptic equations on Riemannian manifolds. *Comm. Anal. Geom.* **10** (4) (2002), 815–846.

Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-4618, U.S.A.
E-mail: mgursky@nd.edu

# Asymptotic solutions for large time of Hamilton–Jacobi equations

Hitoshi Ishii*

**Abstract.** In this article we discuss some recent results on the large-time behavior of solutions of Hamilton–Jacobi equations as well as some ideas and observations behind them and historical remarks concerning them.

**Mathematics Subject Classification (2000).** Primary 35B40; Secondary 35F25, 37J99, 49L25.

**Keywords.** Large-time behavior, asymptotic solutions, Hamilton–Jacobi equations, Aubry sets, weak KAM theory, viscosity solutions.

## 1. Introduction

Hamilton–Jacobi equations play important roles in classical mechanics, geometric optics, optimal control, differential games, etc. We are here interested in global solutions of Hamilton–Jacobi equations. A well-known classical method of finding solutions of Hamilton–Jacobi equations is that of characteristics and its applications have serious difficulties in practice because of developments of shocks in solutions. At the beginning of 1980s M. G. Crandall and P.-L. Lions [10], [11] introduced the notion of viscosity solution in the study of Hamilton–Jacobi equations. It is a notion of generalized solutions for partial differential equations and it is based on the maximum principle while, in this regard, distributions theory is based on integration by parts. This notion has been successfully employed to study fully nonlinear partial differential equations (PDE for short), especially Hamilton–Jacobi equations and fully nonlinear elliptic or parabolic PDE. Important basic features of viscosity solutions are: they enjoy nice properties such as (1) stability under uniform convergence or under the processes of pointwise supremum or infimum, and (2) existence and uniqueness of solutions, under mild assumptions, of boundary value problems or the Cauchy problem for fully nonlinear PDE. See [2], [3], [24], [9] for general overviews and developments of the theory of viscosity solutions.

We recall the definition of viscosity solutions of $F(x, u(x), Du(x)) = 0$ in $\Omega$. Let $u \in C(\Omega, \mathbb{R})$. It is called a viscosity subsolution (resp., supersolution) of $F[u] = 0$ in $\Omega$ if whenever $\varphi \in C^1(\Omega)$ and $u - \varphi$ attains a maximum (resp., minimum) at $y \in \Omega$, then $F(y, u(y), D\varphi(y)) \leq 0$ (resp., $F(y, u(y), D\varphi(y)) \geq 0$). Then, $u \in C(\Omega)$

*The author was supported in part by Grant-in-Aids for Scientific Research, No. 15340051, JSPS.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

is called a viscosity solution of $F[u] = 0$ in $\Omega$ if it is both a viscosity sub- and supersolution of $F[u] = 0$ in $\Omega$. We will be here focused on viscosity solutions, subsolutions, or supersolutions and will suppress the term "viscosity" in what follows as far as there is no danger of confusion.

In this article we consider the stationary Hamilton–Jacobi equation

$$H(x, Du) = 0 \quad \text{in } \Omega, \tag{1}$$

where $\Omega$ is an open subset of $\mathbb{R}^n$, and the Cauchy problem

$$u_t + H(x, Du) = 0 \quad \text{in } \Omega \times (0, \infty), \tag{2}$$

$$u(\cdot, 0) = u_0 \quad \text{in } \Omega. \tag{3}$$

Here $H$ is a continuous function on $\overline{\Omega} \times \mathbb{R}^n$ and $u$ represents the real-valued unknown function on $\Omega$ in the case of (1) or on $\overline{\Omega} \times [0, \infty)$ in the case of (2) and (3), respectively. We write frequently $H[u]$ for $H(x, Du(x))$ for notational simplicity. We will be concerned also with the additive eigenvalue problem

$$H[v] = c \quad \text{in } \Omega. \tag{4}$$

Here the unknown is a pair $(c, v)$ of a constant $c \in \mathbb{R}$ and a function $v$ on $\Omega$ for which $v$ satisfies (4).

The purpose of this article is to review some of recent results concerning the large-time behavior of solutions of (2). An interesting feature of the investigations towards such results is the interaction with the developments of weak KAM theory, and this review will touch upon weak KAM theory. For overviews and developments of weak KAM theory, we refer to [17], [14].

In Section 2 we discuss projected Aubry sets and representation formulas for solutions of (1). In Section 3 the main result concerning the large-time behavior of solutions of (2) are explained. In Section 4 we outline the proof of the main result.

## 2. Projected Aubry sets and representation of solutions

Weak KAM theory introduced by A. Fathi in [15], [17] has changed the viewpoint of uniqueness questions regarding (1).

To begin with, we recall that classical uniqueness or comparison results in viscosity solutions theory applied to the following simple PDE

$$\lambda u + H(x, Du) = 0 \quad \text{in } \Omega, \tag{5}$$

where $\lambda$ is a positive constant, states:

**Theorem 2.1.** *Let $u,\ v \in C(\overline{\Omega})$ be a subsolution and a supersolution of* (5)*, respectively. Assume that $\Omega$ is bounded, that either $u$ or $v$ is locally Lipschitz in $\Omega$, and that $u \le v$ on $\partial\Omega$. Then $u \le v$ in $\Omega$.*

See [2], [3], [8] for general comparison results for Hamilton–Jacobi equations. The above theorem guarantees uniqueness of locally Lipschitz continuous solutions of the Dirichlet problem for (5). We will be concerned mostly with viscosity sub-, super-, or solutions of (1) which are locally Lipschitz continuous, and thus the assumption concerning local Lipschitz continuity of solutions in the above theorem is not any real restriction.

Another way of stating the above theorem is as follows.

**Theorem 2.2.** *Let $u,\ v \in C(\overline{\Omega})$ be solutions, respectively, of $H[u] \leq -\varepsilon$ in $\Omega$ and of $H[v] \geq 0$ in $\Omega$, where $\varepsilon$ is a positive constant. Assume that $\Omega$ is bounded, that either $u$ or $v$ is locally Lipschitz in $\Omega$, and that $u \leq v$ on $\partial\Omega$. Then $u \leq v$ in $\Omega$.*

Let $\Omega = \operatorname{int} B(0, r)$, where $B(0, r)$ denotes the closed ball of radius $r > 0$ with center at the origin and $\operatorname{int} A$ denotes the interior of $A \subset \mathbb{R}^n$. The eikonal equation $|Du| = |x|$ in $\Omega$ has two solutions $u_\pm(x) := \pm(1-|x|^2)/2$, which in addition satisfies the boundary condition, $u(x) = 0$ on $\partial\Omega$. Indeed, the solutions of the Dirichlet problem, $|Du| = |x|$ in $\Omega$ and $u = 0$ on $\partial\Omega$, are given by the family of functions $u_a(x) := \min\{u_+(x), a + u_-(x)\}$ parametrized by $a \in [0,\ 1]$. This example tells us that the Dirichlet problem, $|Du| = |x|$ in $\Omega$ and $u = 0$ on $\partial\Omega$, has many solutions and that the condition, $\lambda > 0$, in Theorem 2.1 is sharp. Thus uniqueness of solution of the Dirichlet problem does not hold in general for (1).

We *assume* henceforth the following two assumptions, the convexity and coercivity of the Hamiltonian $H$:

$$\text{for each } x \in \Omega \text{ the function } p \mapsto H(x, p) \text{ is convex in } \mathbb{R}^n, \tag{6}$$

$$\lim_{r\to\infty} \inf\{H(x, p) \mid x \in \Omega,\ p \in \mathbb{R}^n \setminus B(0, r)\} = \infty. \tag{7}$$

We set $L(x, \xi) = \sup_{p\in\mathbb{R}^n}(\xi\cdot p - H(x, p))$ for $(x, \xi) \in \Omega\times\mathbb{R}^n$, where $\xi\cdot p$ denotes the Euclidean inner product of $\xi,\ p \in \mathbb{R}^n$. We call the function $L\colon \Omega \times \mathbb{R}^n \to (-\infty, \infty]$ the *Lagrangian*.

We define the function $d_H$ on $\Omega \times \Omega$ by

$$d_H(x, y) = \sup\{v(x) \mid H[v] \leq 0,\ v(y) = 0\}.$$

Classical results in viscosity solutions theory assure that the function $d_H$ has the properties:

$$H[d(\,\cdot\,, y)] \leq 0 \quad \text{in } \Omega, \tag{8}$$

$$H[d(\,\cdot\,, y)] \geq 0 \quad \text{in } \Omega \setminus \{y\}. \tag{9}$$

Following [18], we define the (projected) *Aubry set* $\mathcal{A}$ for the Lagrangian $L$ (or for the Hamiltonian $H$) as the subset of $\Omega$ given by

$$\mathcal{A} = \{y \in \Omega \mid H[d(\,\cdot\,, y)] \geq 0 \text{ in } \Omega\}. \tag{10}$$

In view of (8), (9), and (10), it is easily seen that $y \in \Omega \setminus \mathcal{A}$ if and only if $H(y, p) < 0$ for some $p \in D_1^- d_H(y, y)$, where $D_1^- d_H(x, y)$ denotes the subdifferential of $d_H(\cdot, y)$ at $x$. Similarly, we may state that $y \in \Omega \setminus \mathcal{A}$ if and only if there are functions $\sigma \in C(\Omega)$ and $\psi \in C^{0+1}(\Omega)$ such that $\sigma \geq 0$ in $\Omega$, $\sigma(y) > 0$, and $H[\psi] \leq -\sigma$ in $\Omega$.

We now assume for simplicity of presentation that $\Omega$ is an $n$-dimensional torus $\mathbb{T}^n$. The following theorem is an improved version of classical results such as Theorems 2.1 or 2.2

**Theorem 2.3.** *Let $u, v \in C(\Omega)$ be a subsolution and a supersolution of $H = 0$ in $\Omega$, respectively. Assume that $u \leq v$ on $\mathcal{A}$. Then $u \leq v$ in $\Omega$.*

This theorem can be found in [17, Chap. 8]. A key observation for the proof of the above theorem is that for each compact $K \subset \Omega \setminus \mathcal{A}$ there exist a function $\psi_K \in C^{0+1}(\Omega)$ and a constant $\varepsilon_K > 0$ such that $H[\psi_K] \leq -\varepsilon_K$ in a neighborhood $V_K$ of $K$.

Indeed, with such $K$, $\psi_K$, $\varepsilon_K$, and $V_K$, we see that for any $\lambda \in (0, 1)$, the function $u_\lambda := (1 - \lambda)u + \lambda\psi_K$ is a subsolution of $H[u_\lambda] \leq -\lambda\varepsilon_K$ in $V_K$, and hence from Theorem 2.2 that for all $x \in V_K$,

$$(1 - \lambda)u(x) + \lambda\psi_K(x) \leq v(x) + \sup_{\Omega \setminus K}[(1 - \lambda)u + \lambda\psi_K - v],$$

which implies that for all $x \in \Omega \setminus \mathcal{A}$,

$$u(x) \leq v(x) + \sup_{\mathcal{A}}(u - v).$$

In their study of semicontinuous viscosity solutions, E. N. Barron and R. Jensen [4], [5] have observed that under the convexity assumption (6), a function $u \in C(\Omega)$ is a viscosity subsolution of (1) if and only if $H(x, p) \leq 0$ for all $p \in D^- u(x)$ and $x \in \Omega$, where $D^- u(x)$ denotes the subdifferential of $u$ at $x$. A consequence of this observation is that the pointwise infimum of a uniformly bounded family of solutions of (1) yields a solution of (1).

**Theorem 2.4.** *If $u \in C(\Omega)$ is a viscosity solution of $H[u] = 0$ in $\Omega$, then*

$$u(x) = \inf\{d_H(x, y) + u(y) \mid y \in \mathcal{A}\} \quad \textit{for all } x \in \Omega.$$

We remark that if $\mathcal{A} = \emptyset$, then there exists no solution $u$ of $H[u] = 0$ in $\Omega$. The above theorem is a weaker version of [19, Theorem 10.4] which is formulated with the Mather set in place of the Aubry set $\mathcal{A}$ and with quasi-convex Hamiltonian $H$. In the case where $\mathcal{A}$ is a finite set, a corresponding result for the Dirichlet problem for bounded domains has already been obtained in [24].

Another remark here is on the representation of $d_H$ as the value function of an optimal control problem associated with the Hamiltonian $H$. Let $I \subset \mathbb{R}$ be an interval and $\gamma : I \to \Omega$. We say that $\gamma$ is a *curve* if it is absolutely continuous on any compact subinterval of $I$. For $(x, t) \in \Omega \times (0, \infty)$, let $\mathcal{C}(x, t)$ denote the space of curves $\gamma$ on $[0, t]$ such that $\gamma(t) = x$.

**Theorem 2.5.** *Let $x, y \in \Omega$. Then*

$$d_H(x, y) = \inf\left\{ \int_0^t L(\gamma(s), \dot\gamma(s))\mathrm{d}s \;\Big|\; t > 0,\ \gamma \in \mathcal{C}(x, t),\ \gamma(0) = y \right\}, \quad (11)$$

*where $\dot\gamma$ denotes the derivative of $\gamma$.*

Now we turn to the case when $\Omega$ is an open bounded subset of $\mathbb{R}^n$ with regular boundary. We consider the Dirichlet problem

$$H[u] = 0 \quad \text{in } \Omega, \quad (12)$$

$$u|_{\partial\Omega} = g, \quad (13)$$

where $g$ is a given continuous function on $\partial\Omega$. For the Dirichlet problem, we have to modify the definition of the Aubry set and for this we set $\mathcal{A}_D = \mathcal{A} \cup \partial\Omega$, where $\mathcal{A}$ is defined as before. Let $g \in C(\mathcal{A}_D)$ and assume that the following compatibility condition, for the solvability of the Dirichlet problem (12) and (13), holds:

$$g(x) - g(y) \le d_H(x, y) \quad \text{for all } x, y \in \mathcal{A}_D. \quad (14)$$

**Theorem 2.6.** *Under assumption* (14)*, the function $u \in C(\overline{\Omega})$, defined by*

$$u(x) = \inf\{d_H(x, y) + g(y) \mid y \in \mathcal{A}_D\} \quad \textit{for all } x \in \overline{\Omega},$$

*is a solution of* (12) *and* (13)*. Moreover it is a unique solution of* (12) *satisfying $u|_{\mathcal{A}_D} = g$.*

## 3. Asymptotic solutions

The following result is concerned with the unbounded domain $\Omega = \mathbb{R}^n$ and we need a further restriction on $H$. Indeed, we assume that there exist functions $\phi_i$ and $\sigma_i$, with $i = 0, 1$, such that

$$H[\phi_i] \le -\sigma_i \quad \text{in } \mathbb{R}^n,$$

$$\lim_{|x|\to\infty} \sigma_i(x) = \infty,$$

$$\lim_{|x|\to\infty} (\phi_0 - \phi_1)(x) = \infty.$$

Also, we need the following hypothesis:

$$\text{for each } x \in \Omega, \text{ the function } H(x, \cdot) \text{ is strictly convex in } \mathbb{R}^n. \quad (15)$$

We introduce the spaces $\Phi_0$ and $\Psi_0$ of functions on $\mathbb{R}^n$ and on $\mathbb{R}^n \times [0, \infty)$, respectively, as

$$\Phi_0 = \{f \in C(\mathbb{R}^n, \mathbb{R}) \mid \inf_{\mathbb{R}^n}(f - \phi_0) > -\infty\},$$

$$\Psi_0 = \{g \in C(\mathbb{R}^n \times [0, \infty), \mathbb{R}) \mid \inf_{\mathbb{R}^n \times [0,T]}(g - \phi_0) > -\infty, \ \text{ for all } T > 0\}.$$

**Theorem 3.1.** (a) *The additive eigenvalue problem* (4) *has a solution* $(c, v) \in \mathbb{R} \times \Phi_0$*, and moreover the additive eigenvalue* $c$ *is uniquely determined. That is, if* $(d, w) \in \mathbb{R} \times \Phi_0$ *is another solution of* (4)*, then* $d = c$.

(b) *Let* $u_0 \in \Phi_0$. *Then there exists a unique solution* $u \in \Psi_0$ *of the Cauchy problem* (2) *and* (3).

(c) *Let* $\mathcal{A}_c$ *be the Aubry set for the Hamiltonian* $H - c$ *and* $d_{H,c} := d_{H-c}$. *Let* $u \in \Psi_0$ *be the solution of* (2) *and* (3). *Assume that* (15) *holds. Then there exists a solution* $v_0 \in \Phi_0$ *of* (4) *with* $c$ *being the additive eigenvalue for* $H$ *such that*

$$\lim_{t\to\infty} \sup_{x\in B(0,R)} |u(x,t) + ct - v_0(x)| = 0 \quad \textit{for any } R > 0.$$

*Moreover,*

$$v_0(x) = \inf\{d_{H,c}(x, y) + d_{H,c}(y, z) + u_0(z) \mid z \in \mathbb{R}^n,\ y \in \mathcal{A}_c\} \quad \textit{for all } x \in \mathbb{R}^n.$$

The above result is contained in [23]. This result is a variant of those obtained by Fathi, Namah, Roquejoffre, Barles, Souganidis, Davini, Siconolfi, and others for compact manifolds $\Omega$. We refer to [16], [26], [28], [6], [12] for previous results and developments. See also [21] for results in $\mathbb{R}^n$ and [7], [20] for similar results for viscous Hamilton–Jacobi equations. In [21], Hamilton–Jacobi equations of the form $u_t + \alpha x \cdot Du + H_0(Du) = f(x)$ are treated, where $\alpha$ is a positive constant and $H_0, f \in C(\mathbb{R}^n)$. It is assumed that $H_0$ is convex and coercive and that there is a convex function $l \in C(\mathbb{R}^n)$ such that

$$\lim_{|x|\to\infty} (l(-\alpha x) + f(x)) = \infty \quad \text{and} \quad \lim_{|\xi|\to\infty} (L_0 - l)(\xi) = \infty,$$

where $L_0$ is the convex conjugate of $H_0$. If we assume that $H_0$ is strictly convex, then the hypotheses of Theorem 3.1 are satisfied with the choice of $\phi_0(x) := -(1/\sigma)l(-\alpha x)$, $\phi_1(x) := -(1/\alpha)L(-\alpha x)$, $\sigma_0(x) := l(-\alpha x) + f(x) - C$, and $\sigma_1(x) := L(-\alpha x) + f(x)$, where $C$ is a sufficiently large constant.

Another example of $H$ which satisfies the hypotheses of Theorem 3.1 is given by $H(x, p) = H_0(x, p) - f(x)$, where $H_0 \in C(\mathbb{R}^n \times \mathbb{R}^n)$ satisfies (15), (7), and

$$\sup_{\mathbb{R}^n\times B(0,\delta)} |H_0| < \infty \quad \text{for some } \delta > 0,$$

and $f \in C(\mathbb{R}^n)$ satisfies $\lim_{|x|\to\infty} f(x) = \infty$. A possible choice of $\phi_i$, $i = 0, 1$, is as follows: $\phi_0(x) := -(\delta/2)|x|$ and $\phi_1(x) = -\delta|x|$.

## 4. Outline of proof of Theorem 3.1

**4.1. Additive eigenvalue problem.** Additive eigenvalue problem (4) appears in ergodic optimal control or the homogenization of Hamilton–Jacobi equations. In ergodic optimal control the additive eigenvalue $c$ corresponds to averaged long-run optimal costs while $c$ determines the effective Hamiltonian in the homogenization of

Hamilton–Jacobi equations. See [25], [13] for periodic homogenization of Hamilton–Jacobi equations.

To avoid technicalities, we assume in this subsection that $\phi_0 = 0$. One method of solving problem (4) is to approximate it by a regular problem

$$\lambda v_\lambda + H(x, Dv_\lambda) = 0 \quad \text{in } \mathbb{R}^n, \tag{16}$$

where $\lambda$ is a positive constant, and then send $\lambda \to 0$ along an appropriate sequence $\lambda_j \to 0$, to obtain a solution $(c, v) \in \mathbb{R} \times \Phi_0$ of (4) as the limits

$$c := \lim_{j\to\infty} (-\lambda_j v_{\lambda_j}(0)) \quad \text{and} \quad v(x) := \lim_{j\to\infty} (v_{\lambda_j}(x) - v_{\lambda_j}(0)). \tag{17}$$

Indeed, thanks to the coercivity of $H$, we may build a solution $\psi_0 \in C^{0+1}(\mathbb{R}^n)$ of $H[\psi_0] \geq -C_0$ in $\mathbb{R}^n$ for some constant $C_0 > 0$ which satisfies $\phi_0 \leq \psi_0$ in $\mathbb{R}^n$. If $C > 0$ is large enough, then the functions

$$f_\lambda(x) := \phi_0(x) - \lambda^{-1}C \quad \text{and} \quad g_\lambda(x) := \psi_0(x) + \lambda^{-1}C$$

are a subsolution and a supersolution of (16), respectively. The Perron method now yields a solution $v_\lambda \in C^{0+1}(\mathbb{R}^n)$ of (16) which satisfies $f_\lambda \leq v_\lambda \leq g_\lambda$ in $\mathbb{R}^n$. Again the coercivity of $H$ guarantees that the family $\{v_\lambda\}_{\lambda\in(0,1)}$ is locally equi-Lipschitz in $\mathbb{R}^n$, while the inequality $f_\lambda \leq v_\lambda \leq g_\lambda$ in $\mathbb{R}^n$ assures that $\{\lambda v_\lambda(0)\}_{\lambda\in(0,1)}$ is bounded. These observations allow us to pass to the limit in (17).

Another approach to solving (4) is to define the additive eigenvalue $c \in \mathbb{R}$ by

$$c = \inf\{a \in \mathbb{R} \mid \text{there exists a solution } \phi \in C(\mathbb{R}^n) \text{ of } H[\phi] \leq a\}$$

and then to prove that $\mathcal{A}_c \neq \emptyset$. Any pair of $c$ and $v := d_{H,c}(\,\cdot\,, y)$, with $y \in \mathcal{A}_c$, is a solution of (4).

In what follows we *assume* by replacing $H$ by $H - c$, where $c$ is the additive eigenvalue for $H$, that $c = 0$.

**4.2. Critical curves.** An important tool in the weak KAM approach is the collection of critical curves for the Lagrangian $L$. It allows us to analyze the asymptotic behavior for large time of solutions of (2) in the viewpoint of the Lagrangian dynamical system behind (2), which may not be well-defined under our regularity assumptions on $H$.

According to [12], a curve $\gamma : \mathbb{R} \to \mathbb{R}^n$ is said to be *critical* for the Lagrangian $L$ if for any $a, b \in \mathbb{R}$, with $a < b$, and any subsolution $\phi \in C(\mathbb{R}^n)$ of $H[\phi] = 0$ in $\mathbb{R}^n$,

$$\phi(\gamma(b)) - \phi(\gamma(a)) = \int_a^b L(\gamma(s), \dot\gamma(s))\mathrm{d}s.$$

We denote by $\Gamma$ the set of all critical curves $\gamma$. Note that, in general, if $\gamma$ is a curve on $[a, b]$ and $\phi \in C(\mathbb{R}^n)$ is a subsolution of $H[\phi] = 0$ in $\mathbb{R}^n$, then

$$\phi(\gamma(b)) - \phi(\gamma(a)) \leq \int_a^b L(\gamma(s), \dot\gamma(s))\mathrm{d}s. \tag{18}$$

Indeed, we compute by assuming $\phi \in C^1(\mathbb{R}^n)$ that for any curve $\gamma$ on $[a, b]$,

$$\begin{aligned}\phi(\gamma(b)) - \phi(\gamma(a)) &= \int_a^b D\phi(\gamma(s)) \cdot \dot\gamma(s)\mathrm{d}s \\ &\le \int_a^b [L(\gamma, \dot\gamma) + H(\gamma, D\phi(\gamma(s))]\mathrm{d}s \le \int_a^b L(\gamma, \dot\gamma)\mathrm{d}s.\end{aligned}$$

Here we have used the Fenchel inequality: $p \cdot \xi \le H(x, p) + L(x, \xi)$ for all $x$, $p, \xi$ in $\mathbb{R}^n$. The above computation can be justified by the standard mollification technique for general $\phi$ which is locally Lipschitz because of the coercivity of $H$.

**Theorem 4.1.** *For any $y \in \mathcal{A}$ there exists a critical curve $\gamma$ such that $\gamma(0) = y$.*

Existence of critical curves is one of crucial observations in weak KAM theory. See [15], [17], [18], [12] for results on the existence of critical curves.

A main point in the proof of the above theorem is the following general observation concerning the Aubry set, which gives another definition of the Aubry set in terms of the Lagrangian $L$. We remark that this latter definition of the Aubry set has been employed in [15], [18], [19].

**Theorem 4.2.** *Let $y \in \mathbb{R}^n$. Then $y \in \mathcal{A}$ if and only if for any $\varepsilon > 0$,*

$$\inf\left\{ \int_0^t L(\gamma(s), \dot\gamma(s))\mathrm{d}s \;\middle|\; t \ge \varepsilon,\ \gamma \in \mathcal{C}(y, t),\ \gamma(0) = y \right\} = 0.$$

Once we have the above theorem in hand, the proof of Theorem 4.1 goes like this. For any $y \in \mathcal{A}$ and $k \in \mathbb{N}$ we may choose a curve $\gamma_k$ on $[0, T_k]$, where $T_k \ge k$ such that $\gamma_k(0) = \gamma_k(T_k) = y$ and

$$\int_0^{T_k} L(\gamma_k(s), \dot\gamma_k(s))\mathrm{d}s < \frac{1}{k}.$$

We define the curve $\eta_k$ on $[-T_k,\ T_k]$ by setting

$$\eta_k(s) = \begin{cases} \gamma_k(s + T_k) & \text{for } s \in [-T_k, 0], \\ \gamma_k(s) & \text{for } s \in [0, T_k]. \end{cases}$$

Using the observations that

$$\lim_{r\to\infty} \inf\left\{ \frac{L(x, \xi)}{|\xi|} \;\middle|\; x \in B(0, R),\ \xi \in \mathbb{R}^n \setminus B(0, r) \right\} = \infty \quad \text{for any } R > 0, \tag{19}$$

since $H \in C(\mathbb{R}^n \times \mathbb{R}^n)$ and that $\mathcal{A}$ is compact, we may send $k \to \infty$ along a subsequence so that $\eta_k$ has a limit $\gamma$ in $C(\mathbb{R}, \mathbb{R}^n)$, which is a critical curve.

For any $\gamma \in \Gamma$ we have

$$\gamma(t) \in \mathcal{A} \quad \text{for all } t \in \mathbb{R}. \tag{20}$$

This can be seen easily by recalling that for any $y \in \mathbb{R}^n \setminus \mathcal{A}$ there are functions $\phi \in C(\mathbb{R}^n)$ and $\sigma \in C(\mathbb{R}^n)$ such that $\sigma \ge 0$ in $\mathbb{R}^n$, $\sigma(y) > 0$, and $H[\phi] \le -\sigma$ in $\mathbb{R}^n$ and using (18), with $H(x, p)$ replaced by $H(x, p) + \sigma(x)$.

**4.3. Cauchy problem.** To prove existence of a solution of (2) and (3), we may use the well-known formula

$$u(x,t) = \inf\left\{ \int_0^t L(\gamma(s), \dot\gamma(s))\mathrm{d}s + u_0(\gamma(0)) \mid \gamma \in \mathcal{C}(x,t)\right\} \quad \text{for } (x,t) \in \mathbb{R}^n \times (0,\infty). \tag{21}$$

We have to check if this formula really gives a solution of (2) and (3). For this the first thing to do is to check that the function $u$ defined by (21) is a locally bounded function in $\mathbb{R}^n \times (0,\infty)$. Fix a subsolution $\phi \in C(\mathbb{R}^n)$ of $H[\phi] = 0$ in $\mathbb{R}^n$ and set $\phi_2(x) = \min\{\phi(x) - A, \phi_1(x)\}$, where $A > 0$ is a constant. If $A$ is large enough, then $\phi_2$ has the following properties: (a) $H[\phi_2] \le 0$ in $\mathbb{R}^n$ and (b) $\phi_2 \le u_0$ in $\mathbb{R}^n$. Then, for any $(x,t) \in \mathbb{R}^n \times (0,\infty)$ and $\gamma \in \mathcal{C}(x,t)$, we have

$$\phi_2(\gamma(t)) \le u_0(\gamma(0)) + \int_0^t L(\gamma, \dot\gamma)\mathrm{d}s,$$

from which we get $\phi_2(x) \le u(x,t)$.

In order to get a local upper bound of the function $u$ defined by (21), we fix any $(x,t) \in \mathcal{A} \times (0,\infty)$ and choose a curve $\gamma \in \Gamma$ so that $\gamma(t) = x$. Existence of such a critical curve is guaranteed by Theorem 4.1. As before let $\phi \in C(\mathbb{R}^n)$ be a solution of $H[\phi] \le 0$ in $\mathbb{R}^n$. We have

$$\begin{aligned} u(x,t) &\le u_0(\gamma(0)) + \int_0^t L(\gamma, \dot\gamma)\mathrm{d}s \le u_0(\gamma(0)) + \phi(\gamma(t)) - \phi(\gamma(0)) \\ &\le \max_{\mathcal{A}} u_0 + 2\max_{\mathcal{A}} |\phi|. \end{aligned}$$

Since $H \in C(\mathbb{R}^n \times \mathbb{R}^n)$, for each $R > 0$ there are constants $\delta_R > 0$ and $C_R > 0$ such that $L(x,\xi) \le C_R$ for all $(x,\xi) \in B(0,R) \times B(0,\delta_R)$. Fix any $R > 0$ such that $\mathcal{A} \subset B(0,R)$ and any $x \in B(0,R)$. There is a $T_R > 0$, independent of $x$, and a curve (e.g., the line segment connecting a point in $\mathcal{A}$ and $x$) $\gamma_x \in C^1([0,T_R], \mathbb{R}^n)$ such that $\gamma_x(0) \in \mathcal{A}$, $\gamma_x(T_R) = x$, and $|\gamma_x(s)| \le R$, $|\dot\gamma_x(s)| \le \delta_R$ for all $s \in [0,T_R]$. Using the dynamic programming principle which states that for any $x \in \mathbb{R}^n$ and $t, s \in (0,\infty)$,

$$u(x,t+s) = \inf\left\{ \int_0^s L(\gamma(\tau), \dot\gamma(\tau))\mathrm{d}\tau + u(\gamma(0),t) \;\Big|\; \gamma \in \mathcal{C}(x,s)\right\}, \tag{22}$$

we find that

$$\begin{aligned} u(x,t+T_R) &\le \int_0^{T_R} L(\gamma_x(s), \dot\gamma_x(s))\mathrm{d}s + u(\gamma_x(0),t) \\ &\le C_R T_R + \max_{\mathcal{A}} u_0 + 2\max_{\mathcal{A}} |\phi| \quad \text{for all } t \ge 0. \end{aligned}$$

Noting that $u(x,t) \le u_0(x) + tL(x,0)$, we obtain for all $t > 0$,

$$u(x,t) \le \max\{\max_{B(0,R)} (|u_0| + T_R|L(\,\cdot\,,0)|), C_R T_R + \max_{\mathcal{A}} u_0 + 2\max_{\mathcal{A}} |\phi|\}.$$

Setting $u(x, 0) = u_0(x)$ for $x \in \mathbb{R}^n$ and making further standard estimates on $u$, we conclude the following theorem.

**Theorem 4.3.** *The function $u$ belongs to $\Psi_0$ and moreover $u$ is bounded and uniformly continuous on $B(0, R) \times [0, \infty)$ for any $R > 0$.*

Furthermore, using the dynamic programming principle (22), we have:

**Theorem 4.4.** *The function $u$ is a solution of* (2) *and* (3).

Uniqueness of solutions of (2) and (3) follows from the following comparison theorem.

**Theorem 4.5.** *Let $u \in C(\mathbb{R}^n \times [0, \infty))$ and $v \in \Psi_0$ be a subsolution and a supersolution of* (2). *Assume that $u(\cdot, 0) \le v(\cdot, 0)$ in $\mathbb{R}^n$. Then $u \le v$ in $\mathbb{R}^n \times [0, \infty)$.*

An outline of the proof of this theorem goes like this. Let $A$ and $B$ be large positive constants. We set $\psi(x, t) = \phi_1(x) - At$ for $(x, t) \in \mathbb{R}^n \times [0, \infty)$. We may fix $A$ so that $\psi$ is a subsolution of (2). We set $u_B(x, t) = \min\{u(x, t), \psi(x, t) + B\}$ for $(x, t) \in \mathbb{R}^n \times (0, \infty)$ and observe that $u_B$ is a subsolution of (2), that $u_B(\cdot, 0) \le v(\cdot, 0)$ in $\mathbb{R}^n$, and that for each $T > 0$,

$$\lim_{|x| \to \infty} \sup\{(u_B - v)(x, t) \mid t \in [0, T]\} = -\infty. \tag{23}$$

Applying the standard comparison result to $u_B$ and $v$ on the set $B(0, R) \times [0, R)$, with $R > 0$ sufficiently large, we find that $u_B \le v$ in $\mathbb{R}^n \times [0, R)$. Because of the arbitrariness of $R$, $B$, we conclude that $u \le v$ in $\mathbb{R}^n \times [0, \infty)$.

### 4.4. Asymptotic analysis.

**4.4.1. Equilibrium points.** A point $y$ in the Aubry set $\mathcal{A}$ is called an *equilibrium point* if $\min_{p \in \mathbb{R}^n} H(y, p) = 0$ or equivalently $L(y, 0) = 0$. Under the assumption that $\mathcal{A}$ consists only of equilibrium points, the convergence assertion of Theorem 3.1 can be proved in an easy way compared to the general case. To see this, let $u \in \Psi_0$ be the solution of (2) and (3), and set $v_0(x) = \liminf_{t \to \infty} u(x, t)$ for $x \in \mathbb{R}^n$. We then observe in view of the convexity of $H(x, p)$ in $p$ that $v_0$ is a solution of $H[v_0] = 0$ in $\mathbb{R}^n$. Also, we observe at least formally that $u_t \le 0$ in $\mathcal{A} \times (0, \infty)$, which can be stated correctly that the function $t \mapsto u(x, t)$ is nonincreasing in $(0, \infty)$ for any $x \in \mathcal{A}$. Now, by Theorem 4.3 and Dini's lemma, we see that the functions $u(\cdot, t)$ converge to $v_0$ uniformly on $\mathcal{A}$ as $t \to \infty$.

We take a small digression here and state a comparison theorem for (1) with $\Omega = \mathbb{R}^n$, a version of Theorem 2.3 for $\Omega = \mathbb{R}^n$.

**Theorem 4.6.** *Let $u, v \in C(\mathbb{R}^n)$ be a subsolution and a supersolution of $H = 0$ in $\mathbb{R}^n$, respectively. Assume that $u \le v$ on $\mathcal{A}$ and that $v \in \Phi_0$. Then $u \le v$ in $\mathbb{R}^n$.*

For the proof of this theorem, we may assume by a simple modification of $\phi_1$ that $\phi_1$ is a solution of $H[\phi_1] \le 0$ in $\mathbb{R}^n$. We then replace $u$ by $\varepsilon\phi_1 + (1-\varepsilon)u$, with a small $\varepsilon \in (0,1)$, so that we are in the situation that $\lim_{|x|\to\infty}(u-v)(x) = -\infty$, which allows us to work on a ball $B(0,R)$, with sufficiently large $R > 0$. Now, as in the proof of Theorem 2.3, we get $u \le v$ in $\mathbb{R}^n$. This is an outline of the proof of Theorem 4.6.

Back to the main theme, we use the same argument as the proof of Theorem 4.6 just outlined, to control the functions $u(\cdot, t)$ through their restrictions on $\mathcal{A}$ and to conclude the desired convergence of $u(\cdot, t)$ to $v_0$ in $\mathbb{R}^n$ as $t \to \infty$.

We remark that if $\mathcal{A}$ is a finite set, then all the points of $\mathcal{A}$ are equilibrium points. Also, if there is a function $f \in C(\mathbb{R}^n)$ such that $H(x, Df) \le \min_{p\in\mathbb{R}^n} H(x,p)$ for $x \in \mathbb{R}^n$ in the viscosity sense, then all the points of $\mathcal{A}$ are equilibrium points.

**4.4.2. General case.** We turn to the general case. Let $u \in C(\mathbb{R}^n \times [0,\infty))$ be a solution of (2). A formal calculation reveals that for any $\gamma \in \Gamma$, any $t, T \in [0,\infty)$ satisfying $t < T$, and any solution of $\phi \in C(\mathbb{R}^n)$ of $H[\phi] \le 0$ in $\mathbb{R}^n$,

$$\begin{aligned} u(\gamma(T),T) - u(\gamma(t),t) &= \int_t^T \big[Du(\gamma(s),s)\cdot\dot\gamma(s) + u_t(\gamma(s),s)\big]\,\mathrm{d}s \\ &= \int_t^T \big[Du(\gamma(s),s)\cdot\dot\gamma(s) - H(\gamma(s), Du(\gamma(s),s))\big]\,\mathrm{d}s \\ &\le \int_t^T L(\gamma(s),\dot\gamma(s))\mathrm{d}s = \phi(\gamma(T)) - \phi(\gamma(t)). \end{aligned}$$

Indeed, we have

**Lemma 4.7.** *Under the above assumptions, the function $t \mapsto u(\gamma(t),t) - \phi(\gamma(t))$ is nonincreasing on $[0,\infty)$.*

In what follows we denote by $S_t$, with $t \ge 0$, the semigroup generated by (2), i.e., the map $S_t : \Phi_0 \to \Phi_0$ defined by $S_t u_0 = u(\cdot, t)$, where $u$ is the solution in $\Psi_0$ of (2) and (3).

The continuous dependence of the solution of (2) and (3) on the initial data can be stated as follows.

**Theorem 4.8.** *Let $f, g \in \Phi_0$ be a subsolution and a supersolution of $H = 0$ in $\mathbb{R}^n$, respectively. Assume that $f \le g$ in $\mathbb{R}^n$. Then for each $\varepsilon > 0$ there exists $\delta > 0$ such that for any $u_0,\ v_0 \in [f,\ g] \cap C(\mathbb{R}^n)$, if*

$$\max_{B(0,\delta^{-1})} (u_0 - v_0) \le \delta,$$

*then*

$$\sup_{(x,t)\in B(0,\varepsilon^{-1})\times[0,\infty)} (S_t u_0(x) - S_t v_0(x)) \le \varepsilon.$$

*Here* $[f, g]$ *denotes the space of those functions* $w : \mathbb{R}^n \to \mathbb{R}$ *which satisfy* $f \le w \le g$ *in* $\mathbb{R}^n$.

This theorem can be proved by an argument similar to the proof of Theorem 4.5, and we omit presenting it here.

For any $u_0 \in \Phi_0$, the *ω-limit set* $\omega(u_0)$ for the initial point $u_0$ is defined as the set consisting of those $w \in \Phi_0$ for which there exists a sequence $\{t_j\} \subset (0, \infty)$ diverging to infinity such that $S_{t_j} u_0 \to w$ in $C(\mathbb{R}^n)$ as $j \to \infty$. It is obvious from Theorem 4.3 that $\omega(u_0) \neq \emptyset$ for all $u_0 \in \Phi_0$. Another basic property of $\omega$-limit sets, which follows from Theorems 4.8 and 4.3, is the following.

**Lemma 4.9.** *Let* $u_0 \in \Phi_0$ *and let* $\{t_j\}, \{r_j\} \subset (0, \infty)$ *be sequences diverging to infinity such that* $S_{t_j} u_0 \to v$ *and* $S_{t_j + r_j} u_0 \to w$ *in* $C(\mathbb{R}^n)$ *as* $j \to \infty$ *for some* $v, w \in \omega(u_0)$. *Then* $S_{r_j} v \to w$ *in* $C(\mathbb{R}^n)$ *as* $j \to \infty$.

So far, we have needed only the convexity of the Hamiltonian $H(x, p)$ in $p$, but not its strict convexity (15) although this point may not be clear because of the rough presentation. In the next lemma we need the strict convexity assumption (15), which guarantees that $L(x, \xi)$ and $D_\xi L(x, \xi)$ are continuous on the set $\{(x, \xi) \in \mathbb{R}^{2n} \mid L(x, \xi) < \infty\}$. The following lemma is an equivalent of [12, Lemma 5.2] and a key observation for the convergence proof.

**Lemma 4.10.** *Assume that* (15) *holds. Then there exist a* $\delta > 0$ *and a function* $\rho \in C([0, \infty))$, *with* $\rho(0) = 0$, *such that for any* $u_0 \in \Phi_0$, $\gamma \in \Gamma$, $\varepsilon \in (-\delta, \delta)$, *and* $t > 0$,

$$S_t u_0(\gamma(t)) \le u_0(\gamma(\varepsilon t)) + \int_{\varepsilon t}^{t} L(\gamma(s), \dot\gamma(s)) \mathrm{d}s + |\varepsilon t|\, \rho(|\varepsilon|).$$

*Proof of Theorem* 3.1 (c). We denote by $\omega(\Gamma)$ the set of all those curves $\gamma$ to which there correspond a curve $\eta \in \Gamma$ and a sequence $\{t_j\}$ diverging to infinity such that $\eta(\cdot + t_j) \to \gamma$ in $C(\mathbb{R})$ as $j \to \infty$. We set $\mathcal{M} = \{\gamma(0) \mid \gamma \in \omega(\Gamma)\}$. We remark that any $\gamma \in \omega(\Gamma)$ is a critical curve for $L$. Consequently, we have $\mathcal{M} \subset \mathcal{A}$. Moreover, it is easily seen that for any two solutions $\phi, \psi \in C(\mathbb{R}^n)$ of $H = 0$ in $\mathbb{R}^n$, if $\phi \le \psi$ on $\mathcal{M}$, then $\phi \le \psi$ on $\mathcal{A}$.

As before we define the function $v_0 \in \Phi_0$ by

$$v_0(x) = \liminf_{t \to \infty} S_t u_0(x),$$

which is a solution of $H[v_0] = 0$ in $\mathbb{R}^n$.

We prove that $u(\cdot, t) \to v_0$ in $C(\mathbb{R}^n)$ as $t \to \infty$. To this end, it is enough to show that $w = v_0$ for all $w \in \omega(u_0)$. By the definition of $v_0$, we have $v_0 \le w$ in $\mathbb{R}^n$ for all $w \in \omega(u_0)$. Hence, recalling the proof in the case when $\mathcal{A}$ consists only of equilibrium points and using the remark made above, we find that it is enough to show that $w \le v_0$ in $\mathcal{M}$ for all $w \in \omega(u_0)$. (We omit here proving the formula for $v_0$ in the theorem.)

Fix any $w \in \omega(u_0)$ and $y \in \mathcal{M}$. Choose a curve $\gamma \in \Gamma$ and sequences $\{a_j\},\ \{b_j\} \subset (0, \infty)$ diverging to infinity so that, as $j \to \infty$, $\gamma(a_j) \to y$ and $S_{b_j} w \to w$ in $C(\mathbb{R}^n)$. Existence of such a sequence $\{b_j\}$ is assured by Lemma 4.9. We may assume that $c_j := a_j - b_j \to \infty$ as $j \to \infty$. We fix any $s \geq 0$ and apply Lemma 4.10, with $w$ and $\gamma(\cdot + c_j)$ in place of $u_0$ and $\gamma$, respectively, and with $t = b_j$ and $\varepsilon = s/b_j$, to obtain for sufficiently large $j$,

$$S_{b_j} w(\gamma(b_j + c_j)) - w(\gamma(s + c_j)) \leq v_0(\gamma(b_j + c_j)) - v_0(\gamma(s + c_j)) + s\rho(s/b_j),$$

where $\rho \in C([0, \infty))$ is the function from Lemma 4.10. Sending $j \to \infty$ yields

$$w(y) - v_0(y) \leq w(\eta(s)) - v_0(\eta(s)) \quad \text{for all } s \geq 0 \tag{24}$$

and for some $\eta \in \omega(\Gamma)$.

The final step is to show that

$$\liminf_{t \to \infty} [w(\eta(t)) - v_0(\eta(t))] \leq 0, \tag{25}$$

which yields, together with (24), $w(y) \leq v_0(y)$. To do this, we choose sequences $\{t_j\},\ \{\tau_j\} \subset (0, \infty)$ diverging to infinity so that, as $j \to \infty$, $S_{t_j} u_0(\eta(0)) \to v_0(\eta(0))$ and $S_{t_j + \tau_j} u_0 \to w$ in $C(\mathbb{R}^n)$. We calculate by using Lemma 4.7 that

$$\begin{aligned} w(\eta(\tau_j)) - v_0(\eta(\tau_j)) &\leq |w(\eta(\tau_j)) - S_{t_j+\tau_j} u_0(\eta(\tau_j))| + S_{t_j+\tau_j} u_0(\eta(\tau_j)) - v_0(\eta(\tau_j)) \\ &\leq \max_{\mathcal{A}} |w - S_{t_j+\tau_j} u_0| + S_{t_j} u_0(\eta(0)) - v_0(\eta(0)). \end{aligned}$$

Sending $j \to \infty$, we get

$$\limsup_{j \to \infty} \left[ w(\eta(\tau_j)) - v_0(\eta(\tau_j)) \right] \leq 0,$$

which shows that (25) is valid. □

## References


[1] Alvarez, O., Bounded-from-below viscosity solutions of Hamilton-Jacobi equations. *Differential Integral Equations* **10** (3) (1997), 419–436.

[2] Barles, G., *Solutions de viscosité des équations de Hamilton-Jacobi*. Math. Appl. (Berlin) 17, Springer-Verlag, Paris 1994.

[3] Bardi, M., and Capuzzo-Dolcetta, I., *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. With appendices by Maurizio Falcone and Pierpaolo Soravia, Systems Control Found. Appl. Birkhäuser Boston, Inc., Boston, MA, 1997.

[4] Barron, E. N., and Jensen, R., Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians. *Comm. Partial Differential Equations* **15** (12) (1990), 1713–1742.

[5] Barron, E. N., and Jensen, R., Optimal control and semicontinuous viscosity solutions. *Proc. Amer. Math. Soc.* **113** (2) (1991), 397–402.

[6] Barles, G., and Souganidis, P. E., On the large time behavior of solutions of Hamilton-Jacobi equations. *SIAM J. Math. Anal.* **31** (4) (2000), 925–939.

[7] Barles, G., and Souganidis, P. E., Space-time periodic solutions and long-time behavior of solutions to quasi-linear parabolic equations, *SIAM J. Math. Anal.* **32** (6) (2001), 1311–1323.

[8] Crandall, M. G., Ishii, H., and Lions, P.-L., Uniqueness of viscosity solutions of Hamilton-Jacobi equations revisited. *J. Math. Soc. Japan* **39** (4) (1987), 581–596.

[9] Crandall, M. G., Ishii, H., and Lions, P.-L., User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc.* **27** (1992), 1–67.

[10] Crandall, M. G., and Lions, P.-L., Condition d'unicité pour les solutions generalisées des équations de Hamilton-Jacobi du premier ordere. *C. R. Acad. Sci. Paris Sér. I Math.* **292** (1981), 183–186.

[11] Crandall, M. G., and Lions, P.-L., Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.* **277** (1983), 1–42.

[12] Davini, A., and Siconolfi, A., A generalized dynamical approach to the large time behavior of solutions of Hamilton-Jacobi equations. Preprint, 2005.

[13] Evans, L. C., Periodic homogenisation of certain fully nonlinear partial differential equations. *Proc. Roy. Soc. Edinburgh Sect. A* **120** (3–4) (1992), 245–265.

[14] Evans, L. C., A survey of partial differential equations methods in weak KAM theory. *Comm. Pure Appl. Math.* **57** (4) (2004), 445–480.

[15] Fathi, A., Théorème KAM faible et théorie de Mather pour les systèmes lagrangiens. *C. R. Acad. Sci. Paris Sér. I Math.* **324** (9) (1997), 1043–1046.

[16] Fathi, A., Sur la convergence du semi-groupe de Lax-Oleinik. *C. R. Acad. Sci. Paris Sér. I Math.* **327** (3) (1998), 267–270.

[17] Fathi, A., *Weak KAM theorem in Lagrangian dynamics*. To appear.

[18] Fathi, A., and Siconolfi, A., Existence of $C^1$ critical subsolutions of the Hamilton-Jacobi equation. *Invent. Math.* **155** (2) (2004), 363–388.

[19] Fathi, A., and Siconolfi, A., PDE aspects of Aubry-Mather theory for quasiconvex Hamiltonians. *Calc. Var. Partial Differential Equations* **22** (2) (2005), 185–228.

[20] Fujita, Y., Ishii, H., and Loreti, P., Asymptotic solutions of viscous Hamilton-Jacobi equations with Ornstein-Uhlenbeck operator. *Comm. Partial Differential Equations* **31** (6) (2006), 827–848 .

[21] Fujita, Y., Ishii, H., and Loreti, P., Asymptotic solutions of Hamilton-Jacobi equations in Euclidean $n$ space. *Indiana Univ. Math. J.*, to appear.

[22] Ishii, H., A generalization of a theorem of Barron and Jensen and a comparison theorem for lower semicontinuous viscosity solutions. *Proc. Roy. Soc. Edinburgh Sect. A* **131** (1) (2001), 137–154.

[23] Ishii, H., Asymptotic solutions for large time of Hamilton-Jacobi equations in Euclidean $n$ space. Preprint.

[24] Lions, P.-L., *Generalized solutions of Hamilton-Jacobi equations*. Res. Notes in Math. 69, Pitman, Boston, MA, London 1982.

[25] Lions, P.-L., Papanicolaou, G., and Varadhan, S., Homogenization of Hamilton-Jacobi equations. Unpublished preprint.

[26] Namah, G., and Roquejoffre, J.-M., Remarks on the long time behaviour of the solutions of Hamilton-Jacobi equations. *Comm. Partial Differential Equations* **24** (5–6) (1999), 883–893.

[27] Rockafellar, T., *Convex Analysis*. Princeton Math. Ser. 28, Princeton University Press, Princeton, 1970.

[28] Roquejoffre, J.-M., Convergence to steady states or periodic solutions in a class of Hamilton-Jacobi equations. *J. Math. Pures Appl.* (9) **80** (1) (2001), 85–104.

Department of Mathematics, Faculty of Education and Integrated Arts and Sciences,
Waseda University, Nishi-waseda, Shinjuku-ku, Tokyo, 169-8050 Japan
E-mail: ishii@edu.waseda.ac.jp

# The weak-coupling limit of large classical and quantum systems

Mario Pulvirenti*

**Abstract.** In this contribution we illustrate the delicate transition from the microscopic description of a particle system, given in terms of fundamental equations as the Newton or the Schrödinger equation, to the reduced kinetic picture, given in terms of the Boltzmann and Landau equations which are obtained under suitable scaling limits. Special emphasis is given to the so called weak-coupling limit.

The content of the lecture is mostly devoted to the very many open problems, rather than to the few known results.

**Mathematics Subject Classification (2000).** Primary 35Q99; Secondary 82C40.

**Keywords.** Boltzmann equation, Landau equation, weak-coupling limit.

## 1. The Boltzmann and Landau equations

The present section is largely discursive: its scope is the heuristic introduction of the Boltzmann and Landau equations on the basis of physical arguments.

In 1872 Ludwig Boltzmann, starting from the mathematical model of elastic balls and using mechanical and statistical considerations, established an evolution equation to describe the behavior of a rarefied gas. The starting point of the Boltzmann analysis is to renounce to study the behavior of a gas in terms of the detailed motion of the molecules which constitute it because of their huge number. It is rather better to investigate a function $f(x, v)$ which is the probability density of a given particle, where $x$ and $v$ denote position and velocity of such a particle. Actually $f(x, v)dxdv$ is often confused with the fraction of molecules falling in the cell of the pase space of size $dx\,dv$ around $x$, $v$. The two concept are not exactly the same but they are asymptotically equivalent (when the number of particles is diverging) if a law of large numbers holds.

The Boltzmann equation is the following:

$$(\partial_t + v \cdot \nabla_x) f = Q(f, f) \tag{1.1}$$

*Most of the considerations developed here are due to a systematic collaboration of the author with D. Benedetto, F. Castella and R. Esposito.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

where $Q$, the collision operator, is defined for $\lambda > 0$ by

$$Q(f, f) = \lambda^{-1} \int dv_1 \int_{S_+} dn\, (v - v_1) \cdot n\, [f(x, v') f(x, v_1') - f(x, v) f(x, v_1)] \tag{1.2}$$

and

$$v' = v - n[n \cdot (v - v_1)], \quad v_1' = v_1 + n[n \cdot (v - v_1)]. \tag{1.3}$$

Also $n$ (the impact parameter) is a unitary vector and $S_+ = \{n \mid n \cdot (v - v_1) \geq 0\}$. Note that $v'$, $v_1'$ are the outgoing velocities after a collision of two elastic balls with incoming velocities $v$ and $v_1$ and centers $x$ and $x + dn$, being $d$ the diameter of the spheres. Obviously the collision takes place if $n \cdot (v - v_1) \geq 0$. Equations (1.3) are consequence of the energy, momentum and angular momentum conservation. Note also that $d$ does not enter in equation (1.1) as a parameter.

As fundamental features of equation (1.1) we have the conservation in time of the following five quantities

$$\iint dx dv f(x, v; t) v^\alpha \tag{1.4}$$

with $\alpha = 0, 1, 2$ expressing conservation of the probability, momentum and energy.

Moreover Boltzmann introduced the (kinetic) entropy defined by

$$H(f) = \int dx \int dv f \log f(x, v) \tag{1.5}$$

and proved the famous H-theorem asserting the decreasing of $H(f(t))$ along the solutions to equation (1.1).

Finally, in case of bounded domains, the distribution defined for $\beta > 0$:

$$M(v) = \text{const}\, e^{-\beta v^2},$$

called Maxwellian distribution, is stationary for the evolution given by equation (1.1). In addition $M$ minimizes $H$ among all distributions with zero mean velocity, and given energy.

In conclusion Boltzmann was able to introduce an evolutionary equation with the remarkable properties of expressing mass, momentum, energy conservations, but also the trend to the thermal equilibrium. In other words he tried to conciliate the Newton laws with the second principle of Thermodynamics.

Boltzmann's heuristic argument in deriving equation (1.1) is, roughly speaking, the following. The molecular system we are considering consists of $N$ identical particles of diameter $d$ in the whole space $\mathbb{R}^3$ and we denote by $x_1, v_1, \ldots, x_N, v_N$ a state of the system, where $x_i$ and $v_i$ indicate the position and the velocity of the particle $i$. The particles cannot overlap, that is the centers of two particles cannot be at distance smaller than the diameter $d$.

The particles are moving freely up to the first contact instant, that is the first time in which two particles arrive at distance $d$. Then the pair interacts performing

an elastic collision. This means that they change instantaneously their velocities, according to the conservation of the energy, linear and angular momentum. After the first collision the system goes on by iterating the procedure. Here we neglect triple collisions because unlikely. The evolution equation for a tagged particle is of the form

$$(\partial_t + v \cdot \nabla_x) f = \text{Coll} \tag{1.6}$$

where Coll denotes the variation of $f$ due to the collisions. We have

$$\text{Coll} = G - L \tag{1.7}$$

where $L$ and $G$ (loss and gain term respectively) are the negative and positive contribution to the variation of $f$ due to the collisions. More precisely $Ldxdvdt$ is the probability of our test particle to disappear from the cell $dxdv$ of the phase space because of a collision in the time interval $(t, t + dt)$ and $Gdxdvdt$ is the probability to appear in the same time interval for the same reason. Let us now consider the sphere of center $x$ with radius $d$ and a point $x + dn$ over the surface, where $n$ denotes the generic unit vector. Consider also the cylinder with base area $dS = d^2 dn$ and height $|V|dt$ along the direction of $V = v_2 - v$.

Then a given particle (say particle 2) with velocity $v_2$, can contribute to $L$ because it can collide with our test particle in the time $dt$, provided it is localized in the cylinder and if $V \cdot n \leq 0$. Therefore the contribution to $L$ due to the particle 2 is the probability of finding such a particle in the cylinder (conditioned to the presence of the first particle in $x$). This quantity is $f_2(x, v, x + nd, v_2)|(v_2 - v) \cdot n|d^2 dn dv_2 dt$, where $f_2$ is the joint distribution of two particles. Integrating in $dn$ and $dv_2$ we obtain that the total contribution to $L$ due to any predetermined particle is:

$$d^2 \int dv_2 \int_{S_-} dn f_2(x, v, x + nd, v_2)|(v_2 - v) \cdot n| \tag{1.8}$$

where $S_-$ is the unit hemisphere $(v_2 - v) \cdot n < 0$. Finally we obtain the total contribution multiplying by the total number of particles:

$$L = (N - 1)d^2 \int dv_2 \int_{S_-} dn f_2(x, v, x + nd, v_2)|(v_2 - v) \cdot n|. \tag{1.9}$$

The gain term can be derived analogously by considering that we are looking at particles which have velocities $v$ and $v_2$ after the collisions so that we have to integrate over the hemisphere $S_+ = (v_2 - v) \cdot n > 0$:

$$G = (N - 1)d^2 \int dv_2 \int_{S_+} dn f_2(x, v, x + nd, v_2)|(v_2 - v) \cdot n|. \tag{1.10}$$

Summing $G$ and $-L$ we get

$$\text{Coll} = (N - 1)d^2 \int dv_2 \int dn f_2(x, v, x + nd, v_2)(v_2 - v) \cdot n. \tag{1.11}$$

which, however, is not a very useful expression because the time derivative of $f$ is expressed in term of another object namely $f_2$. An evolution equation for $f_2$ will imply $f_3$, the joint distribution of three particles and so on up to arrive to the total particle number $N$. Here the basic Boltzmann's main assumption enters, namely that two given particles are uncorrelated if the gas is rarefied, namely:

$$f(x, v, x_2, v_2) = f(x, v) f(x_2, v_2). \tag{1.12}$$

Condition (1.12), called *propagation of chaos*, seems contradictory at a first sight: if two particles collide, correlations are created. Even though we could assume equation (1.12) at some time, if the test particle collides with the particle 2, such an equation cannot be satisfied anymore after the collision.

Before discussing the propagation of chaos hypothesis, we first analyze the size of the collision operator. We remark that, in practical situations for a rarefied gas, the combination $Nd^3 \approx 10^{-4}\text{cm}^3$ (that is the volume occupied by the particles) is very small, while $Nd^2 = O(1)$. This implies that $G = O(1)$. Therefore, since we are dealing with a very large number of particles we are tempted to perform the limit $N \to \infty$ and $d \to 0$ in such a way that $d^2 = O(N^{-1})$. As a consequence the probability that two tagged particles collide (which is of the order of the surface of a ball, that is $O(d^2)$), is negligible. However the probability that a given particle performs a collision with any one of the remaining $N-1$ particles (which is $O(Nd^2) = O(1)$) is not negligible. Condition (1.12) is referring to two preselected particles (say particle 1 and particle 2) so that it is not unreasonable to conceive that it holds in the limiting situation in which we are working.

However we cannot insert (1.12) in (1.11) because this latter equation refer both to instants before and after the collision and, if we know that a collision took place, we certainly cannot invoke (1.12). Hence we assume (1.12) in the loss term and work over the gain term to keep advantage of the factorization property which will be assumed *only* before the collision.

Coming back to equation (1.10) for the outgoing pair velocities $v, v_2$ (satisfying the condition $(v_2 - v) \cdot n > 0$) we make use of the continuity property

$$f_2(x, v, x + nd, v_2) = f_2(x, v', x + nd, v_2') \tag{1.13}$$

where the pair $v', v_2'$ is pre-collisional. On $f_2$ expressed before the collision we can reasonably apply condition (1.12) obtaining:

$$\begin{aligned} G - L = (N-1)r^2 \int dv_2 \int_{S_-} dn (v - v_2) \cdot n \\ \cdot [f(x, v') f(x - nd, v_2') - f(x, v) f(x + nd, v_2)] \end{aligned} \tag{1.14}$$

after a change $n \to -n$ in the Gain term. This transforms the pair $v', v_2'$ from a pre-collisional to a post-collisional pair.

Finally, in the limit $N \to \infty$, $r \to 0$, $Nd^2 = \lambda^{-1}$ we find equation (1.1) where $Q$, the collision operator, has the form (1.2). The parameter $\lambda$, called *mean*

*free path*, represents, roughly speaking, the typical length a particle can cover without undergoing any collision.

Equation (1.1) is has a statistical nature and it is not equivalent to the Hamiltonian dynamics from which it has been derived. In particular, due to the H-Theorem, it is not time reversal.

The heuristic arguments we have developed so far can be extended to different potentials than that of the hard-sphere systems. If the particles interact via a two-body interaction $\phi = \phi(x)$ the resulting Boltzmann equation is equation (1.1) with

$$Q(f, f) = \int dv_1 \int_{S_+} dn B(v - v_1; n) [f' f_1' - f f_1], \tag{1.15}$$

where we are using the usual short hand notation:

$$f' = f(x, v'), \quad f_1' = f(x, v_1'), \quad f = f(x, v), \quad f_1 = f(x, v_1) \tag{1.16}$$

and $B = B(v - v_1; n)$ is a suitable function of the relative velocity and the impact parameter, proportional to the cross-section relative to the potential $\phi$. Another equivalent, some times convenient way to express $Q$ is

$$Q(f, f) = \int dv_1 \int dv' \int dv_1' W(v, v_1 | v', v_1') [f' f_1' - f f_1] \tag{1.17}$$

with

$$W(v, v_1 | v', v_1') = w(v, v_1 | v', v_1') \cdot \delta(v + v_1 - v' + v_1') \delta\left(\frac{1}{2}(v^2 + v_1^2 - (v')^2 + (v_1')^2)\right). \tag{1.18}$$

and $w$ a suitable kernel. All the qualitative properties as the conservation laws and the H-theorem are obviously still valid.

The arguments we have used in deriving the Boltzmann equation are delicate and require a more rigorous and deeper analysis. If we want that the Boltzmann equation is not a phenomenological model, derived by assumptions *ad hoc* and justified by its practical relevance, but rather a consequence of a mechanical model, we must derive it rigorously from a logical and mathematical viewpoint. In particular the propagation of chaos should be not an hypothesis but the statement of a theorem.

Many scientists, among them Loschmidt, Zermelo and Poincaré, outlined inconsistencies between the irreversibility of the equation and the reversible character of the Hamiltonian dynamics. Boltzmann argued the statistical nature of his equation and his answer to the irreversibility paradox was that *most* of the configurations behave as expected by the thermodynamical laws. However he did not have the probabilistic tools for formulating in a precise way the statements of which he had a precise intuition. In 1949 H. Grad [25] stated clearly the limit $N \to \infty$, $d \to 0$, $Nd^2 \to$ const, where $N$ is the number of particles and $d$ is the diameter of the molecules, in which the Boltzmann equation is expected to hold. This limit is usually called the Boltzmann-Grad (or low-density limit).

The problem of a rigorous derivation of the Boltzmann equation was an open and challenging problem for a long time. O. E. Lanford [31] showed that, although for a very short time, the Boltzmann equation can be derived starting from the mechanical model of the hard-sphere system. The proof has a deep content but is relatively simple from a technical view point. Later on [28] it has been proved that this technique can be adapted to prove a result holding globally in time, but for the special situation of a rare claud of gas expanding in the vacuum.

We address the reader to references [31], [14] and [19] for a deeper discussion on the validation problem of the Boltzmann equation. We also warmly suggest the monograph [13] for a critical and historical discussion on the Boltzmann equation and the scientist who conceived it.

A preliminary problem to the validation of the Boltzmann equation for an arbitrary time interval, still open in general, is the construction of a global solution, hopefully unique. See [14] and [42] for the state of art of existence problems at present times. We just mention that the most general result we have up to now is due to Di Perna and Lions [16] who showed the existence of suitable weak solutions to equation (1.1). However we still do not know whether such solutions, which preserve mass, momentum and satisfy the H-theorem, are unique and preserve also the energy.

The Boltzmann equation works for rarefied gas, however one can ask whether a useful kinetic picture can be invoked for dense gas. Here we want to describe a situation in which the gas particles are weakly interacting, but $N = O(r^{-3})$ being $r \ll 1$ the interaction length of the particles. To express the weakness of the interaction, we assume that the two-particle potential is $O(\sqrt{r})$. In this case we want to compute the total momentum variation for a unit time. Note that the force is $O\left(\frac{1}{\sqrt{r}}\right)$ but acts on the time interval $O(r)$. The momentum variation due to the single scattering is therefore $O(\sqrt{r})$. The number of particles met by a test particles is $O\left(\frac{1}{r}\right)$. Hence the total momentum variation for unit time is $O\left(\frac{1}{\sqrt{r}}\right)$. However this variation, in case of homogeneous gas and symmetric force, should be zero in the average. If we compute the variance, it should be $\frac{1}{r}O(\sqrt{r})^2 = O(1)$. As a consequence of this central limit type of argument we expect that the kinetic equation which holds in the limit (if any), should be a diffusion equation in velocity variable.

A more convincing argument will be presented in the next section. For the moment let us now argue at level of kinetic equation. Consider the collision operator in the form (1.17). Suppose that $\varepsilon > 0$ is a small parameter. To express the fact that the transferred momentum is small, we rescale $w$ as $\frac{1}{\varepsilon^3} w\left(\frac{p}{\varepsilon}\right)$. In addition we also rescale the mean-free path inverse by a factor $\frac{1}{\varepsilon}$ to take into account the high density situation. The collision operator becomes:

$$Q_\varepsilon(f, f) = \frac{1}{\varepsilon^4} \int dv_1 \int dp w\left(\frac{p}{\varepsilon}\right) \delta(p^2 + (v - v_1) \cdot p)[f' f_1' - f f_1] \qquad (1.19)$$

$$
\begin{aligned}
&= \frac{1}{2\pi\varepsilon^2}\int dv_1 \int dp\ w(p) \int_{-\infty}^{+\infty} ds e^{is(p^2\varepsilon+(v-v_1)\cdot p)} \\
&\quad \cdot [f(v+\varepsilon p) f(v_1-\varepsilon p) - f(v) f(v_1)] \\
&= \frac{1}{2\pi\varepsilon}\int dv_1 \int dp\ w(p) \int_0^1 d\lambda \int_{-\infty}^{+\infty} ds e^{is(p^2\varepsilon+(v-v_1)\cdot p)} \\
&\quad \cdot p(\nabla_v - \nabla_{v_1}) f(v+\varepsilon\lambda p) f(v_1-\varepsilon\lambda p).
\end{aligned}
$$

Here the smooth function $w$, which modulates the collision, is assumed depending only on $p$ trough its modulus. The $\delta$ appearing in equation (1.19) expresses the energy conservation.

To outline the behavior of $Q_\varepsilon(f, f)$ in the limit $\varepsilon \to 0$, we introduce a test function $\varphi$ for which, after a change of variables ( here $(\cdot,\cdot)$ denotes the scalar product in $L_2(v)$):

$$
\begin{aligned}
(\varphi, Q_\varepsilon(f,f)) &= \frac{1}{2\pi\varepsilon}\int dv \int dv_1 \int dp\ w(p) \int_0^1 d\lambda \int_{-\infty}^{+\infty} ds e^{is(p^2(\varepsilon-2\varepsilon\lambda)+(v-v_1)\cdot p)} \\
&\quad \cdot \varphi(v-\varepsilon\lambda p)\ p\cdot(\nabla_v - \nabla_{v_1}) f f_1 \\
&= \frac{1}{2\pi\varepsilon}\int dv \int dv_1 \int dp w(p) \int_0^1 d\lambda \int_{-\infty}^{+\infty} ds e^{is(v-v_1)\cdot p} \\
&\quad \cdot [\varphi(v) + \varepsilon p\cdot\nabla_v\varphi(v)] p\cdot(\nabla_v - \nabla_{v_1}) f f_1 \\
&\quad + \frac{1}{2\pi}\int dv \int dv_1 \int dp\ w(p) \int_{-\infty}^{+\infty} ds e^{is(v-v_1)\cdot p)}\varphi(v) \\
&\quad \cdot isp^2 \int_0^1 d\lambda(1-2\lambda)\ p\cdot(\nabla_v - \nabla_{v_1}) f f_1 + O(\varepsilon).
\end{aligned}
\tag{1.20}
$$

Note now that the term $O(\varepsilon^{-1})$ vanishes because of the symmetry $p \to -p$ ($w$ is even). Also the imaginary part of the $O(1)$ term is vanishing, being null the integral in $d\lambda$. As a result:

$$
\begin{aligned}
(\varphi, Q_\varepsilon(f,f)) = \frac{1}{2\pi}\int dv \int dv_1 \int dp\ w(p) \int_{-\infty}^{+\infty} ds\ e^{is(v-v_1)\cdot p)} \\
\cdot p\cdot\nabla_v\varphi\ p\cdot(\nabla_v - \nabla_{v_1}) f f_1 + O(\varepsilon).
\end{aligned}
\tag{1.21}
$$

Therefore we have recovered the kinetic equation (1.1) with a new collision operator

$$
Q_L(f,f) = \int dv_1 \nabla_v\, a(\nabla_v - \nabla_{v_1}) f f_1, \tag{1.22}
$$

where $a = a(v - v_1)$ denotes the matrix

$$
a_{i,j}(V) = \int dp\ w(p)\ \delta(V\cdot p)\ p_i p_j. \tag{1.23}
$$

This matrix can be handled in a better way by introducing polar coordinates:

$$a_{i,j}(V) = \frac{1}{|V|}\int dp\ |p|w(p)\ \delta(\hat{V}\cdot\hat{p})\ \hat{p}_i\hat{p}_j \cdot \frac{B}{|V|}\int d\hat{p}\ \delta(\hat{V}\cdot\hat{p})\ \hat{p}_i\hat{p}_j, \tag{1.24}$$

where $\hat{V}$ and $\hat{P}$ are the versor of $V$ and $p$ respectively and

$$B = \int_0^{+\infty} drr^3 w(r). \tag{1.25}$$

Note that $B$ is the only parameter describing the interaction appearing in the equation. Finally a straightforward computation yields:

$$a_{i,j}(V) = \frac{B}{|V|}(\delta_{i,j} - \hat{V}_i\hat{V}_j). \tag{1.26}$$

The collision operator $Q_L$ has been introduced by Landau ([32]) for the study of a weakly interacting dense plasma. Note that the qualitative properties of the solutions to the Landau equation are the same as for the Boltzmann equation as regards the basic conservation laws and the H-theorem.

A rigorous derivation of the Landau equation starting from the Boltzmann equation in the grazing collision limit (that is what we presented here at a formal level), has been obtained in [1], [24] and [43] for spatially homogeneous solutions. The (diverging) asymptotics for the Coulomb forces is discussed in [15] and in [43]. However, in the present lecture, we are interested in deriving the Landau equation in terms of particle systems. In the next section we present a formal derivation outlining the difficulties in trying a rigorous proof.

## 2. Weak-coupling limit for classical systems

We consider a classical system of $N$ identical particles of unitary mass. Positions and velocities are denoted by $q_1, \ldots, q_N$ and $v_1, \ldots, v_N$. The Newton equations reads as:

$$\frac{d}{d\tau}q_i = v_i, \quad \frac{d}{d\tau}v_i = \sum_{\substack{j=1,\ldots,N:\\ j\neq i}} F(q_i - q_j). \tag{2.1}$$

Here $F = -\nabla\phi$ denotes the interparticle (conservative) force, $\phi$ the two-body interaction potential and $\tau$ the time.

We are interested in a situation where the number of particles $N$ is very large and the interaction quite moderate. In addition we look for a reduced or macroscopic description of the system. Namely if $q$ and $\tau$ refer to the system seen in a microscopic

scale, we introduce $\varepsilon > 0$ a small parameter expressing the ratio between the macro and microscales. Indeed it is often convenient to rescale equation (2.1) in terms of the macroscopic variables

$$x = \varepsilon q, \quad t = \varepsilon \tau$$

whenever the physical variables of interest are varying on such scales and are almost constant on the microscopic scales. Therefore, rescaling the potential according to

$$\phi \to \sqrt{\varepsilon}\phi, \tag{2.2}$$

system (2.1), in terms of the $(x, t)$ variables, becomes

$$\frac{d}{dt}q_i = v_i, \quad \frac{d}{dt}v_i = -\frac{1}{\sqrt{\varepsilon}} \sum_{\substack{j=1,\dots,N: \\ j\neq i}} \nabla\phi\Big(\frac{x_i - x_j}{\varepsilon}\Big). \tag{2.3}$$

Note that the velocities are automatically unscaled. Moreover we also assume that $N = O(\varepsilon^{-3})$, namely the density is $O(1)$.

Let $W^N = W^N(X_N, V_N)$ be a probability distribution on the phase space of the system. Here $(X_N, V_N)$ denote the set of positions and velocities:

$$X_N = x_1, \dots, x_N, \quad V_N = v_1, \dots, v_N.$$

Then from equations (2.3) we obtain the following Liouville equation:

$$(\partial_t + V_N \cdot \nabla_N) W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}} \big(T_N^\varepsilon W^N\big)(X_N, V_N), \tag{2.4}$$

where $V_N \cdot \nabla_N = \sum_{i=1}^N v_i \cdot \nabla_{x_i}$ and $(\partial_t + V_N \cdot \nabla_N)$ is the usual free stream operator. Also, we have introduced the operator

$$(T_N^\varepsilon W^N)(X_N, V_N) = \sum_{0<k<\ell\leq N} (T_{k,\ell}^\varepsilon W^N)(X_N, V_N), \tag{2.5}$$

with

$$T_{k,\ell}^\varepsilon W^N = \nabla\phi\Big(\frac{x_k - x_\ell}{\varepsilon}\Big) \cdot (\nabla_{v_k} - \nabla_{v_\ell}) W^N. \tag{2.6}$$

To investigate the limit $\varepsilon \to 0$ it is convenient to introduce the BBKGY hierarchy for the $j$- particle distributions defined as

$$\begin{aligned} f_j^N(X_j, V_j) = \int dx_{j+1} \dots \int dx_N \int dv_{j+1} \dots \int dv_N \\ \cdot W^N(X_j, x_{j+1}, \dots, x_N; V_j, v_{j+1}, \dots, v_N) \end{aligned} \tag{2.7}$$

for $j = 1, \dots, N-1$. Obviously, we set $f_N^N = W^N$. Note that BBKGY stands for Bogoliubov, Born, Green, Kirkwood and Yvon, the names of physicists who introduced independently this system of equations (see e.g. [2] and [19]).

From now on we shall suppose that, due to the fact that the particles are identical, the objects which we have introduced ($W^N$, $f_j^N$) are all symmetric in the exchange of particles.

A partial integration of the Liouville equation (2.4) and standard manipulations give us the following hierarchy of equations (for $1 \le j \le N$):

$$\Big(\partial_t + \sum_{k=1}^{j} v_k \cdot \nabla_k\Big) f_j^N = \frac{1}{\sqrt{\varepsilon}} T_j^\varepsilon f_j^N + \frac{N-j}{\sqrt{\varepsilon}} C_{j+1}^\varepsilon f_{j+1}^N. \tag{2.8}$$

The operator $C_{j+1}^\varepsilon$ is defined as

$$C_{j+1}^\varepsilon = \sum_{k=1}^{j} C_{k,j+1}^\varepsilon, \tag{2.9}$$

and

$$\begin{aligned} &C_{k,j+1}^\varepsilon f_{j+1}(x_1, \dots, x_j; v_1, \dots, v_j) \\ &\quad = -\int dx_{j+1} \int dv_{j+1} F\Big(\frac{x_k - x_\ell}{\varepsilon}\Big) \nabla_{v_k} f_{j+1}(x_1, x_2, \dots, x_{j+1}; v_1, \dots, v_{j+1}). \end{aligned} \tag{2.10}$$

$C_{k,j+1}^\varepsilon$ describes the "collision" of particle $k$, belonging to the $j$-particle subsystem, with a particle outside the subsystem, conventionally denoted by the number $j+1$ (this numbering uses the fact that all particles are identical). The total operator $C_{j+1}^\varepsilon$ takes into account all such collisions. The dynamics of the $j$-particle subsystem is governed by three effects: the free-stream operator, the collisions "inside" the subsystem (the $T$ term), and the collisions with particles "outside" the subsystem (the $C$ term).

We finally fix the initial value $\{f_j^0\}_{j=1}^N$ of the solution $\{f_j^N(t)\}_{j=1}^N$ assuming that $\{f_j^0\}_{j=1}^N$ is factorized, that is, for all $j = 1, \dots, N$

$$f_j^0 = f_0^{\otimes j}, \tag{2.11}$$

where $f_0$ is a given one-particle distribution function. This means that the states of any pair of particles are statistically uncorrelated at time zero. Of course such a statistical independence is destroyed at time $t > 0$. Dynamics creates correlations and equation (2.8) shows that the time evolution of $f_1^N$ is determined by the knowledge of $f_2^N$ which turns out to be dependent on $f_3^N$ and so on. However, since the interaction between two given particle is going to vanish in the limit $\varepsilon \to 0$, we can hope that such statistical independence is recovered in the same limit. Note that the physical meaning of the propagation of chaos here is quite different from that arising in the contest of the Boltzmann equation. Here two particles can interact but the effect of the collision is small, while in a low-density regime the effect of a collision between two given particles is large but quite unlikely.

Therefore we expect that in the limit $\varepsilon \to 0$ the one-particle distribution function $f_1^N$ converges to the solution of a suitable nonlinear kinetic equation $f$ which we are going to investigate.

If we expand $f_j^N(t)$ as a perturbation of the free flow $S(t)$ defined as

$$(S(t)f_j)(X_j, V_j) = f_j(X_j - V_j t, V_j), \tag{2.12}$$

we find

$$\begin{aligned} f_j^N(t) = S(t)f_j^0 + \frac{N-j}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) C_{j+1}^{\varepsilon} f_{j+1}^N(t_1) dt_1 \\ + \frac{1}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) T_j^{\varepsilon} f_j^N(t_1) dt_1. \end{aligned} \tag{2.13}$$

We now try to keep information on the limit behavior of $f_j^N(t)$. Assuming for the moment that the time evolved $j$-particle distributions $f_j^N(t)$ are smooth (in the sense that the derivatives are uniformly bounded in $\varepsilon$), then

$$\begin{aligned} &C_{j+1}^{\varepsilon} f_{j+1}^N(X_j; V_j; t_1) \\ &\quad = -\varepsilon^3 \sum_{k=1}^{j} \int dr \int dv_{j+1} F(r) \cdot \nabla_{v_k} f_{j+1}(X_j, x_k - \varepsilon r; V_j, v_{j+1}, t_1). \end{aligned} \tag{2.14}$$

Assuming now, quite reasonably, that

$$\int dr F(r) = 0, \tag{2.15}$$

we find that

$$C_{j+1}^{\varepsilon} f_{j+1}^N(X_j; V_j; t_1) = O(\varepsilon^4)$$

provided that $D_v^2 f_{j+1}^N$ is uniformly bounded. Since

$$\frac{N-j}{\sqrt{\varepsilon}} = O\big(\varepsilon^{7/2}\big)$$

we see that the second term in the right-hand side of (2.13) does not give any contribution in the limit.

Moreover

$$\begin{aligned} &\int_0^t S(t-t_1) T_j^{\varepsilon} f_j^N(t_1) dt_1 \\ &\quad = \sum_{i \neq k} \int_0^t dt_1 F\Big( \frac{(x_i - x_k) - (v_i - v_k)(t-t_1)}{\varepsilon} \Big) g(X_j, V_j; t_1) \end{aligned} \tag{2.16}$$

where $g$ is a smooth function.

Obviously the above time integral is $O(\varepsilon)$ so that also the last term in the right-hand side of (2.13) does not give any contribution in the limit. Then we are facing the alternative: either the limit is trivial or the time evolved distributions are not smooth. This is indeed a bed new because, if we believe that the limit is not trivial (actually we expect to get the Landau equation, according to the previous discussion) a rigorous proof of this fact seems problematic.

The difficulty in obtaining a-priori estimates induce us to exploit the full series expansion of the solution, namely

$$f_1^N(t) = \sum_{n\geq 0}\sum_{G_n} K(G_n)\int_0^t dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n \cdot \left[S(t-t_1)O_1 S(t_1-t_2)\ldots O_n S(t_n)\right] f_m^0. \tag{2.17}$$

Here $O_j$ is either an operator $C$ or $T$ expressing a creation of a new particle or a recollision between two particles respectively. $G_n$ is a graph namely a sequence of indices

$$(r_1, l_1),\ (r_2, l_2), \ldots, (r_n, l_n)$$

where $(r_j, l_j)$, $r_j < l_j$ is the pair of indices of the particles involved in the interaction at time $t_j$. $m-1$ is the number of particles created in the process. It is convenient to represent the generic graph in the following way.

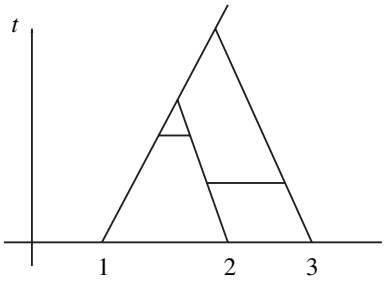



Figure 1

Here the legs of the graph denotes the particles and the nodes the creation of new particles (operators $C$). Recollisions (operators $T$) are represented by horizontal links. For instance the graph in the figure is

$$(1,2),\quad (1,3),\quad (1,3),\quad (2,3)$$

($m=3$), and the integrand in equation (2.17) in this case is

$$\left[S(t-t_1)C_{1,2}S(t_1-t_2)C_{1,3}S(t_2-t_3)T_{1,3}S(t_3-t_4)T_{2,3}S(t_4)\right] f_3^0. \tag{2.18}$$

Note that the knowledge of the graph determines completely the sequence of operators in the right-hand side of (2.17). Finally the factor $K(G_n)$ takes into account

the divergences:

$$K(G_n) = O\left(\left(\frac{1}{\sqrt{\varepsilon}}\right)^n \varepsilon^{-3(m-1)}\right). \tag{2.19}$$

We are not able to analyze the asymptotic behaviour of each term of the expansion (2.17) however we can compute the limit for $\varepsilon \to 0$ of the few terms up to the second order (in time). We have:

$$\begin{aligned} &g^N(x_1, v_1; t) \\ &\quad = f^0(x_1 - v_1 t, v_1) + \frac{N-1}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) C^{\varepsilon}_{1,2} S(t_1) f_2^0 dt_1 \\ &\qquad + \frac{(N-1)}{\varepsilon} \frac{(N-2)}{\varepsilon} \sum_{j=1,2} \int_0^t \int_0^{t_1} dt_2 S(t-t_1) C^{\varepsilon}_{1,2} S(t_1 - t_2) C^{\varepsilon}_{j,3} S(t_2) f_3^0 \\ &\qquad + \frac{N-1}{\varepsilon} \int_0^t dt_1 \int_0^{t_1} dt_2 S(t-t_1) C^{\varepsilon}_{1,2} S(t_1 - t_2) T^{\varepsilon}_{1,2} S(t_2) f_2^0. \end{aligned} \tag{2.20}$$

Here the right-hand side of (2.20) defines $g^N$.

The second and third term in (2.20) corresponding to the graphs in Figure 2 are indeed vanishing as follows by the use of the previous arguments. The most interesting term is the last one (collision–recollision) in Figure 3.

Figure 2

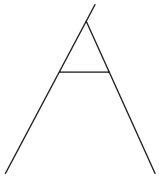


Figure 3

To handle this term we denote by $w = v_1 - v_2$ the relative velocity and note that, for

a given function $u$,

$$\begin{aligned} &S(t_1-t_2)T^{\varepsilon}_{1,2}u(x_1,x_2;v_1,v_2)\\ &\quad=-F\Big(\frac{(x_1-x_2)-w(t_1-t_2)}{\varepsilon}\Big)\\ &\qquad\cdot[(\nabla_{v_1}-\nabla_{v_2})u](x_1-v_1(t_1-t_2),x_2-v_2(t_1-t_2);v_1,v_2) \qquad (2.21)\\ &\quad=-F\Big(\frac{(x_1-x_2)-w(t-t_1)}{\varepsilon}\Big)\\ &\qquad\cdot(\nabla_{v_1}-\nabla_{v_2}+(t_1-t_2)(\nabla_{x_1}-\nabla_{x_2}))S(t_1-t_2)u(x_1,x_2;v_1,v_2) \end{aligned}$$

Therefore the last term in the right-hand side of (2.21) is

$$\begin{aligned} &\frac{N-1}{\varepsilon}\int_0^t dt_1 S(t-t_1)\int_0^{t_1}dt_2\int dx_2\int dv_2\\ &\quad\cdot F\Big(\frac{x_1-x_2}{\varepsilon}\Big)\cdot\nabla_{v_1}F\Big(\frac{(x_1-x_2)-w(t_1-t_2)}{\varepsilon}\Big) \qquad (2.22)\\ &\qquad\cdot(\nabla_{v_1}-\nabla_{v_2}+(t_1-t_2)(\nabla_{x_1}-\nabla_{x_2}))S(t_1)f_2^0(x_1,x_2;v_1,v_2). \end{aligned}$$

Setting now $r=\frac{x_1-x_2}{\varepsilon}$ and $s=\frac{t_1-t_2}{\varepsilon}$ then

$$\begin{aligned} &g^N(x_1,v_1)\\ &\quad=(N-1)\varepsilon^3\int_0^t dt_1\int_0^{\frac{t_1}{\varepsilon}}ds\int dr\int dv_2\ F(r)\cdot\nabla_{v_1}\cdot F(r-ws) \qquad (2.23)\\ &\qquad\cdot(\nabla_{v_1}-\nabla_{v_2}+\varepsilon s(\nabla_{x_1}-\nabla_{x_2}))S(t_1-\varepsilon s)f_2^0(x_1,x_2;v_1,v_2)+O(\sqrt{\varepsilon}). \end{aligned}$$

The formal limit is of (2.20) is

$$g(t)=S(t)f_0+\int_0^t dt_1 S(t-t_1)\nabla_{v_1}a(v_1-v_2)\ (\nabla_{v_1}-\nabla_{v_2})\ S(t_1)f_2^0, \qquad (2.24)$$

where (using $F(r)=-F(-r)$) the matrix $a$ is given by

$$\begin{aligned} a(w)&=\int dr\int_0^{+\infty}ds\ F(r)\otimes F(r-ws)\\ &=\frac{1}{2}\int dr\int_{-\infty}^{+\infty}ds\ F(r)\otimes F(r-ws)\\ &=\frac{1}{2}\left(\frac{1}{2\pi}\right)^3\int_{-\infty}^{+\infty}ds\int dk k\otimes k\ \hat{\phi}(k)^2e^{i(w\cdot k)s}\\ &=\left(\frac{1}{8\pi}\right)^2\int dk k\otimes k\ \hat{\phi}(k)^2\delta(w\cdot k). \end{aligned} \qquad (2.25)$$

Here the interaction potential $\phi$ has been assumed spherically symmetric. Therefore the matrix $a$ has the same form (1.26) with $B$ given by

$$B=\left(\frac{1}{8\pi}\right)^2\int_0^{+\infty}drr^3\hat{\phi}(r)^2. \qquad (2.26)$$

Consider now the Landau equation

$$(\partial_t + v \cdot \nabla_x) f = Q_L(f, f) \tag{2.27}$$

with the collision operator $Q_L$ given by (1.22) and the matrix $a$ given by

$$a_{i,j}(V) = \frac{B}{|V|}(\delta_{i,j} - \hat{V}_i \hat{V}_j), \tag{2.28}$$

$B$ being defined by (2.26). We obtain the following (infinite) hierarchy of equations

$$(\partial_t + V_j \cdot \nabla_{X_j}) f_j = C_{j+1} f_{j+1} \tag{2.29}$$

for the quantities

$$f_j(t) = f(t)^{\otimes j}, \tag{2.30}$$

where $f(t)$ solves equation (2.27). Accordingly $C_{j+1} = \sum_k C_{k,j+1}$, where

$$C_{k,j+1} f_{j+1}(x_1, \ldots, x_j; v_1, \ldots, v_j) = \prod_{r \neq k} f(x_r.v_r) Q_L(f, f)(x_k, v_k). \tag{2.31}$$

Therefore $f$ has the following series expansion representation

$$\begin{aligned} f(t) = \sum_{n \geq 0} \int_0^t dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n \\ \cdot \left[ S(t - t_1) C_2 S(t_1 - t_2) C_3 \ldots C_n S(t_n) \right] f_{n+1}^0. \end{aligned} \tag{2.32}$$

As matter of fact we showed the formal convergence of $g^N$ to the first two terms of the expansion (2.32), namely we have an agreement between the particle system (2.17) and the solution to the Landau equation (2.32) at least up to the first order in time (or second order in the potential). Although the above arguments can be made rigorous under suitable assumption on the initial condition $f_0$ and the potential $\phi$, it seems difficult to show the convergence of the whole series. On the other hand it is clear that the graphs which should contribute in the limit are those formed by a collision–recollision sequence like:

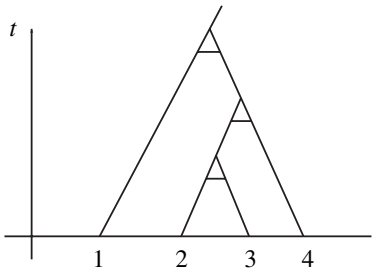



Figure 4

For those terms it is probably possible to show the convergence. For instance the case in the figure has the asymptotics

$$\int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 \big[S(t-t_1)C_{1,2}S(t_1-t_2)C_{2,3}S(t_2-t_3)C_{3,4}S(t_3)\big] f_4^0. \quad (2.33)$$

However the proof that all other graphs are vanishing in the limit is not easy. Even more difficult is a uniform control of the series expansion (2.17), even for short times. As we shall see in the next section, something more can be obtained for quantum systems under the same scaling limit. Note finally that the diffusion coefficient found here given by equation (2.26), is different from that obtained in the grazing collision limit (see (1.25)). Indeed the transition kernel $w$ appearing in equation (1.18) is in general different from $\hat{\phi}^2$. Actually to recover the Landau equation by a low-density limit (to get the Boltzmann equation) and then selecting the grazing collision part, is not equivalent to the direct, and more physical, weak-coupling limit.

## 3. Weak-coupling limit for quantum systems

We consider now the quantum analog of the system considered in Section 2, namely $N$ identical quantum particles with unitary mass in $\mathbb{R}^3$.

The interaction between particles is still a two-body potential $\phi$ so that the total potential energy is taken as

$$U(x_1, \dots, x_N) = \sum_{i<j} \phi(x_i - x_j). \quad (3.1)$$

The associated Schrödinger equation reads

$$i\partial_t \Psi(X_N, t) = -\frac{1}{2}\Delta_N \Psi(X_N, t) + U(X_N)\Psi(X_N, t), \quad (3.2)$$

where $\Delta_N = \sum_{i=1}^N \Delta_i$, $\Delta_i$ is the Laplacian with respect to the $x_i$ variables, $X_N = x_1, \dots, x_N$ and $\hbar$ is normalized to unity.

As for the classical system considered in Section 2 we rescale the equation and the potential by

$$x \to \varepsilon x, \quad t \to \varepsilon t, \quad \phi \to \sqrt{\varepsilon}\phi. \quad (3.3)$$

The resulting equation is,

$$i\varepsilon\partial_t \Psi^\varepsilon(X_N, t) = -\frac{\varepsilon^2}{2}\Delta_N \Psi^\varepsilon(X_N, t) + U_\varepsilon(X_N)\, \Psi^\varepsilon(X_N, t), \quad (3.4)$$

where

$$U_\varepsilon(x_1, \dots, x_N) = \sqrt{\varepsilon} \sum_{i<j} \phi\Big(\frac{x_i - x_j}{\varepsilon}\Big). \quad (3.5)$$

We want to analyze the limit $\varepsilon \to 0$ in the above equations, when $N = \varepsilon^{-3}$.

Note that this limit looks, at a first sight, similar to a semiclassical (or high frequence) limit. It is not so: indeed the potential varies on the same scale of the typical oscillations of the wave functions so that the scattering process is a genuine quantum process. Obviously, due to the oscillations, we do not expect that the wave function does converge to something in the limit. The right quantity to look at was introduced by Wigner in 1922 [44] to deal with kinetic problems. It is called the Wigner transform (of $\Psi^\varepsilon$) and is defined as

$$W^N(X_N, V_N) = \left(\frac{1}{2\pi}\right)^{3N} \int dY_N \; e^{iY_N \cdot V_N} \overline{\Psi}^\varepsilon \left(X_N + \frac{\varepsilon}{2} Y_N\right) \Psi^\varepsilon \left(X_N - \frac{\varepsilon}{2} Y_N\right). \quad (3.6)$$

As it is standard, $W^N$ satisfies a transport-like equation, completely equivalent to the Schrödinger equation:

$$(\partial_t + V_N \cdot \nabla_N) W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}} \big(T_N^\varepsilon W^N\big)(X_N, V_N). \quad (3.7)$$

The operator $T_N^\varepsilon$ on the right-hand-side of (3.7) plays the same role of the classical operator denoted with the same symbol in Section 2. It is

$$(T_N^\varepsilon W^N)(X_N, V_N) = \sum_{0<k<\ell\leq N} (T_{k,\ell}^\varepsilon W^N)(X_N, V_N), \quad (3.8)$$

where each $T_{k,\ell}^\varepsilon$ describes the interaction of particle $k$ with particle $\ell$:

$$\begin{aligned}(T_{k,\ell}^\varepsilon W^N)&(X_N, V_N) \\ &= \frac{1}{i} \left(\frac{1}{2\pi}\right)^{3N} \int dY_N \, dV'_N \; \mathrm{e}^{iY_N \cdot (V_N - V'_N)} \\ &\quad \cdot \left[\phi\left(\frac{x_k - x_\ell}{\varepsilon} - \frac{y_k - y_\ell}{2})\right) - \phi\left(\frac{x_k - x_\ell}{\varepsilon} + \frac{y_k - y_\ell}{2}\right)\right] W^N(X_N, V'_N).\end{aligned} \quad (3.9)$$

Equivalently, we may write

$$\begin{aligned}(T_{k,\ell}^\varepsilon W^N)&(X_N, V_N) \\ &= -i \sum_{\sigma=\pm 1} \sigma \int \frac{dh}{(2\pi)^3} \; \hat{\phi}(h) \; \mathrm{e}^{i\frac{h}{\varepsilon}(x_k - x_\ell)} \\ &\quad \cdot W^N\big(x_1, \ldots, x_N; v_1, \ldots, v_k - \sigma\tfrac{h}{2}, \ldots, v_\ell + \sigma\tfrac{h}{2}, \ldots, v_N\big).\end{aligned} \quad (3.10)$$

Note that $T_{k,\ell}^\varepsilon$ is a pseudodifferential operator which formally converge, at fixed $\varepsilon$, for $\hbar \to 0$ (here $\hbar = 1$) to its classical analog. Note also that in (3.10), "collisions" may take place between *distant* particles ($x_k \neq x_\ell$). However, such distant collisions are penalized by the highly oscillatory factor $\exp(ih(x_k - x_\ell)/\varepsilon)$. These oscillations

turn out to play a crucial role throughout the analysis, and they explain why collisions tend to happen when $x_k = x_\ell$ in the limit $\varepsilon \to 0$.

The formalism we have introduced is similar to the one of the classical case so that we proceed as before by transforming equation (3.7) into a hierarchy of equations. We introduce the partial traces of the Wigner transform $W^N$, denoted by $f_j^N$. They are defined through the following formula, valid for $j = 1, \dots, N-1$:

$$f_j^N(X_j, V_j) = \int dx_{j+1} \dots \int dx_N \int dv_{j+1} \dots \int dv_N \cdot W^N(X_j, x_{j+1}, \dots, x_N; V_j, v_{j+1}, \dots, v_N). \tag{3.11}$$

Obviously, we set $f_N^N = W^N$. The function $f_j^N$ is the kinetic object that describes the state of the $j$ particles subsystem at time $t$.

Due to the fact that the particles are identical, the wave function $\Psi$, as well as $W^N$ and $f_j^N$, are assumed to be symmetric in the exchange of particle, a property that is preserved in time.

Proceeding then as in the derivation of the BBKGY hierarchy for classical systems, we readily transform equation (3.7) into the following hierarchy:

$$\Big(\partial_t + \sum_{k=1}^{j} v_k \cdot \nabla_k\Big) f_j^N(X_j, V_j) = \frac{1}{\sqrt{\varepsilon}} T_j^\varepsilon f_j^N + \frac{N-j}{\sqrt{\varepsilon}} C_{j+1}^\varepsilon f_{j+1}^N, \tag{3.12}$$

where

$$C_{j+1}^\varepsilon = \sum_{k=1}^{j} C_{k,j+1}^\varepsilon, \tag{3.13}$$

and $C_{k,j+1}^\varepsilon$ is defined by

$$\begin{aligned} &C_{k,j+1}^\varepsilon f_{j+1}^N(X_j; V_j) \\ &\quad = -i \sum_{\sigma=\pm 1} \sigma \int \frac{dh}{(2\pi)^3} \int dx_{j+1} \int dv_{j+1}\, \hat{\phi}(h)\, e^{i\frac{h}{\varepsilon}(x_k - x_{j+1})} \\ &\qquad \cdot f_{j+1}^N\big(x_1, x_2, \dots, x_{j+1}; v_1, \dots, v_k - \sigma\tfrac{h}{2}, \dots, v_{j+1} + \sigma\tfrac{h}{2}\big). \end{aligned} \tag{3.14}$$

As before the initial value $\{f_j^0\}_{j=1}^N$ is assumed completely factorized: for all $j = 1, \dots, N$, we suppose

$$f_j^0 = f_0^{\otimes j}, \tag{3.15}$$

where $f_0$ is a one-particle Wigner function, and $f^0$ is assumed to be a probability distribution.

In the limit $\varepsilon \to 0$, we expect that the $j$-particle distribution function $f_j^N(t)$, that solves the hierarchy (3.12) with initial data (3.15), tends to be factorized for all times: $f_j^N(t) \sim f(t)^{\otimes j}$ (propagation of chaos).

As for the classical case, if $f_{j+1}$ is smooth, then

$$\begin{aligned} &C^{\varepsilon}_{k,j+1} f^N_{j+1}(X_j; V_j) \\ &\quad = -i\varepsilon^3 \sum_{\sigma=\pm 1} \sigma \int \frac{dh}{(2\pi)^3} \hat{\phi}(h) \int dr \int dv_{j+1}\, e^{ih\cdot r} \\ &\qquad \cdot f^N_{j+1}\big(X_j, x_k - \varepsilon r;\ v_1, \dots, v_k - \sigma\tfrac{h}{2}, \dots, v_{j+1} + \sigma\tfrac{h}{2}\big) = O(\varepsilon^4). \end{aligned} \tag{3.16}$$

Indeed, setting $\varepsilon = 0$ in the integrand, the integration over $r$ produces $\delta(h)$. As a consequence the integrand is independent of $\sigma$ and the sum vanishes. Therefore the integral is $O(\varepsilon)$. Also

$$\begin{aligned} &\frac{1}{\sqrt{\varepsilon}} \int_0^t dt_1 S(t-t_1) T_{r.k} f^N_j(t_1) \\ &\quad = -i \sum_{\sigma=\pm 1} \sigma \int_0^t dt_1 \frac{dh}{(2\pi)^3} \hat{\phi}(h) \\ &\qquad \cdot e^{i\frac{h}{\varepsilon}\cdot(x_r - x_k) - (v_r - v_k)(t-t_1)} f^N_j(X_j - V_j(t-t_1); V_j; t_1) \end{aligned} \tag{3.17}$$

is weakly vanishing, by a stationary phase argument (see [4]). Therefore we are in the same situation as for the classical case for which we are led to analyze the asymptotics of $g^N$ (see (2.20)) which means to study the limit of the collision–recollision term:

$$\frac{N-1}{\varepsilon} \int_0^t dt_1 \int_0^{t_1} d\tau_1\ S(t-t_1) C_{1,2} S(t_1 - \tau_1) T_{1.2} S(\tau_1) f^0_2. \tag{3.18}$$

Explicitly it looks as follows:

$$\begin{aligned} &-\frac{N-1}{\varepsilon} \sum_{\sigma,\sigma'=\pm 1} \sigma\sigma' \int_0^t dt_1 \int_0^{t_1} d\tau_1 \int dx_2 \int dv_2 \int \frac{dh}{(2\pi)^3} \int \frac{dk}{(2\pi)^3} \\ &\quad \cdot \hat{\phi}(h)\ \hat{\phi}(k)\ \mathrm{e}^{i\frac{h}{\varepsilon}\cdot(x_1 - x_2 - v_1(t-t_1))} \mathrm{e}^{i\frac{k}{\varepsilon}\cdot(x_1 - x_2 - v_1(t-t_1) - (v_1 - v_2 - \sigma h)(t_1 - \tau_1))} \\ &\qquad \cdot f^0_2\left(x_1 - v_1 t + \sigma\tfrac{h}{2} t_1 + \sigma'\tfrac{k}{2}\tau_1, x_2 - v_2 t_1 - \sigma\tfrac{h}{2} t_1 - \sigma'\tfrac{k}{2}\tau_1;\right. \\ &\qquad\quad \left. \cdot\ v_1 - \sigma\tfrac{h}{2} - \sigma'\tfrac{k}{2}, v_2 + \sigma\tfrac{h}{2} + \sigma'\tfrac{k}{2}\right). \end{aligned} \tag{3.19}$$

By the change of variables

$$t_1 - \tau_1 = \varepsilon s_1 \ (\text{i.e. } \tau_1 = t_1 - \varepsilon s_1), \quad \xi = (h+k)/\varepsilon, \tag{3.20}$$

we have

$$\begin{aligned} (3.19) = -(N-1)\,\varepsilon^3 &\sum_{\sigma,\sigma'=\pm 1} \sigma\sigma' \int_0^t dt_1 \int_0^{t_1/\varepsilon} ds_1 \int dx_2 \int dv_2 \int \frac{d\xi}{(2\pi)^3} \int \frac{dk}{(2\pi)^3} \\ &\cdot \hat{\phi}(-k + \varepsilon\xi_1)\ \hat{\phi}(k)\ \mathrm{e}^{i\xi\cdot\left(x_1 - x_2 - v_1(t-t_1)\right)} \mathrm{e}^{-is_1 k\cdot(v_1 - v_2 - \sigma(-k+\varepsilon\xi))} f^0_2(\dots), \end{aligned}$$

In the limit $\varepsilon \to 0$, the above formula gives the asymptotics

$$(3.19)\quad \underset{\varepsilon\to 0}{\sim} -\sum_{\sigma,\sigma'=\pm 1} \sigma\sigma' \int_0^t dt_1 \int dv_2 \int \frac{dk}{(2\pi)^3} \tag{3.21}$$
$$|\hat{\phi}(k)|^2 \left(\int_0^{+\infty} e^{-is_1 k\cdot(v_1-v_2+\sigma k)}\, ds_1\right) f_2^0\left(x_1 - v_1 t - (\sigma-\sigma')\tfrac{k}{2}t_1,\right.$$
$$\left. x_1 - v_1(t-t_1) - v_2 t_1 + (\sigma-\sigma')\tfrac{k}{2}t_1;\, v_1 + (\sigma-\sigma')\tfrac{k}{2},\, v_2 - (\sigma-\sigma')\tfrac{k}{2}\right).$$

In [4], we completely justify formula (3.21) and its forthcoming consequences.

Now, we turn to identifying the limiting value obtained in (3.21). To do so, we observe that symmetry arguments allow us to replace the integral in $s$ by its real part:

$$\mathrm{Re}\int_0^{\infty} e^{-is_1 k\cdot(v_1-v_2+\sigma k)}\, ds_1 = \pi\delta(k\cdot(v_1-v_2+\sigma k)). \tag{3.22}$$

Using formula (3.22) we realize that the contribution $\sigma = -\sigma'$ in (3.21) gives rise to the gain term:

$$\int_0^t dt_1 \int dv_2 \int d\omega\, B(\omega, v_1-v_2) \cdot f_2^0(x_1 - v_1(t-t_1) - v_1' t_1, x_2 - v_2(t-t_1) - v_2' t_1;\, v_1', v_2'), \tag{3.23}$$

where

$$B(\omega, v) = \frac{1}{8\pi^2}|\omega\cdot v|\,|\hat{\phi}(\omega\,(\omega\cdot v))|^2. \tag{3.24}$$

Similarly, the term $\sigma = \sigma'$ in (2.2) yields the loss term:

$$\int_0^t dt_1 \int dv_2 \int d\omega\, B(\omega, v_1-v_2) f_2^0(x_1 - v_1 t, x_2 - v_2(t-t);\, v_1, v_2). \tag{3.25}$$

By the same arguments used in the previous section we can conclude that the full series expansion (2.17) (of course for the present quantum case) agrees, up to the second order in the potential, with

$$S(t)f_0 + \int_0^t dt_1 S(t-t_1) Q(S(t_1)f_0, S(t_1)f_0) \tag{3.26}$$

where

$$Q(f,f) = \int dv_1 \int d\omega\, B(\omega, v-v_1)[f'f_1' - ff_1] \tag{3.27}$$
$$= \int dv_1 \int dh |\hat{\phi}(h)|^2 \delta((h\cdot(v-v_1+h))[f(v+h)f(v_1-h) - f(v)f(v_1)].$$

In other words the kinetic equation which comes out is the Boltzmann equation with cross-section $B$.

We note once more that the $\delta$ function in equation (3.27) expresses the energy conservation, while the momentum conservation is automatically satisfied.

Note that the cross-section $B$ is the only quantum factor in the purely classical expression (3.27). It retains the quantum features of the elementary "collisions".

An important comment is in order. Why is the kinetic equation for quantum systems of Boltzmann type in contrast with the classical case where we got a diffusion? The answer is related to the asymptotics of a single scattering (see [35], [36] and [8]). For quantum systems the probability of a zero angle scattering is finite (that is a sort of tunnel effect), while for a classical particle we have surely a small deviation from the free motion. Therefore a quantum particle, in this asymptotic regime, is going to perform a jump process (in velocity) rather than a diffusion.

From a mathematical view point we observe that [4] proves more than agreement up to second order. We indeed consider the subseries (of the full series expansion expressing $f_j^N(t)$) formed by *all* the collision–recollision terms. In other words, we consider the subseries of $f_j^N(t)$ given by

$$\sum_{n\geq 1} \sum_{\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n} \varepsilon^{-4n} \int_0^t dt_1 \int_0^{t_1} d\tau_1 \; S(t-t_1) C^{\varepsilon}_{\alpha_1,\beta_1} S(t_1-\tau_1) T^{\varepsilon}_{\alpha_1,\beta_1} \cdots \int_0^{\tau_{n-1}} dt_n \int_0^{t_n} d\tau_n \; S(\tau_{n-1}-t_n) C^{\varepsilon}_{\alpha_n,\beta_n} S(t_n-\tau_n) T^{\varepsilon}_{\alpha_n,\beta_n} S(\tau_n) f^0_{j+n+1}. \tag{3.28}$$

Here the sum runs over all possible choices of the particles number $\alpha$'s and $\beta$'s, namely we sum over the subset of graphs of the form in Figure 4. We establish in [4] that the subseries (3.28) is indeed absolutely convergent, for short times, uniformly in $\varepsilon$. Moreover, we prove that it approaches the corresponding complete series expansion obtained by solving iteratively the Boltzmann equation with collision operator given by equation (3.27) extending and making rigorous the above argument.

However, this does not completely finishes the proof yet: the true series expansion of $f_j^N(t)$ contains many more terms than those we consider in (3.28) and we are not able to show uniform bound on the full series. Thus a mathematical justification of the quantum Boltzmann equation is a still an open and difficult problem. More recently we proved in [7], although under severe assumptions on the potential, that all other terms than those considered in the subseries (3.28) are vanishing in the limit, but this is, unfortunately, not yet conclusive.

## 4. The weak coupling limit in the Bose–Einstein or the Fermi–Dirac statistics

From a physical viewpoint it is certainly more realistic to consider particles obeying the Fermi–Dirac or Bose–Einstein statistics, than considering the Maxwell–Boltzmann situation. In this case, the starting point still is the rescaled Schrödinger equation (3.4),

or the equivalent hierarchy (3.12). The only new point is that we cannot take a totally uncorrelated initial datum as in (3.15). Indeed, the Fermi–Dirac or Bose–Einstein statistics yield correlations even at time zero. In this perspective, the most uncorrelated states one can introduce, and that do not violate the Fermi–Dirac or Bose–Einstein statistics, are the so-called quasi-free states. They have, in terms of the Wigner formalism, the following form:

$$f_j(x_1, v_1, \dots, x_j, v_j) = \sum_{\pi \in \mathcal{P}_j} \theta^{s(\pi)} \, f_j^{\pi}(x_1, v_1, \dots, x_j, v_j), \tag{4.1}$$

where each $f_j^{\pi}$ has the value

$$f_j^{\pi}(x_1, v_1, \dots, x_j, v_j) = \int dy_1 \dots dy_j \int dw_1 \dots dw_j \; e^{i(y_1 \cdot v_1 + \dots + y_j \cdot v_j)} \tag{4.2}$$

$$\prod_{k=1}^{j} e^{-\frac{i}{\varepsilon} w_k \cdot (x_k - x_{\pi(k)})} \; e^{-\frac{i}{2} w_k \cdot (y_k + y_{\pi(k)})} \; f\Big( \frac{x_k + x_{\pi(k)}}{2} + \varepsilon \frac{y_k - y_{\pi(k)}}{4}, w_k \Big)$$

and $f$ is a given one-particle Wigner function. Here $\mathcal{P}_j$ denotes the group of all the permutations of $j$ objects and $\pi$ its generic element.

Note that the Maxwell–Boltzmann case treated so far is recovered by the contribution due the permutation $\pi =$ identity.

Note also that quasi-free states converge weakly to the completely factorized states as $\varepsilon \to 0$, that is a physically obvious fact because the quantum statistics become irrelevant in the semiclassical limit. However the dynamics take place on the scale $\varepsilon$ so that the effects of the statistics are present in the limit. Indeed it is expected that the one-particle distribution function $f_1^N(t)$ converges to the solution of the following cubic Boltzmann equation:

$$(\partial_t + v \cdot \nabla_x) \, f(t, x, v) = Q_{\mathrm{w},\theta}(f, f, f)(t, x, v), \tag{4.3}$$

$$\begin{aligned} Q_\theta(f, f, f)(t, x, v) = \int dv_1 \, d\omega \; B_\theta(\omega, v - v_1) \\ \cdot \big[ f(x, v') f(x, v_1')(1 + 8\pi^3 \theta f(x, v) \, f(x, v_1)) \\ - f(x, v) f(x, v_1)(1 + 8\pi^3 \theta f(x, v') f(x, v_1')) \big]. \end{aligned} \tag{4.4}$$

Here $\theta = +1$ or $\theta = -1$, for the Bose–Einstein or the Fermi–Dirac statistics respectively. Finally, $B_\theta$ is the symmetrized or antisymmetrized cross-section derived from $B$ (see (3.24)) in a natural way.

As we see, the modification of the statistics transforms the quadratic Boltzmann equation of the Maxwell–Boltzmann case, into a cubic one (fourth order terms cancel). Also, the statistics affects the form of the cross-section and $B$ has to be (anti)symmetrized into $B_\theta$. The collision operator (4.4) has been introduced by Uehling and Uhlembeck in 1933 on the basis of purely phenomenological considerations [41].

Plugging in the hierarchy (3.12) an initial datum satisfying (4.1), we can follow the same procedure as for the Maxwell–Boltzmann statistics: we write the full perturbative series expansion expressing $f_j^N(t)$ in terms of the initial datum and try to analyse its asymptotic behaviour.

As we did before, we first restrict our attention to those terms of degree less than two in the potential.

The analysis up to second order is performed in [5]. We actually recover here equation (4.3), (4.4) with the suitable $B_\theta$. Now the number of terms to control is much larger due to the sum over all permutations that enters the definition (4.1) of the initial state. Also, the asymptotics is much more delicate. In particular, we stress the fact that the initial datum brings its own highly oscillatory factors in the process, contrary to the Maxwell–Boltzmann case where the initial datum is uniformly smooth, and where the oscillatory factors simply come from the collision operators $T$ and $C$. In [5] we consider the second order graphs

Figure 5

which, because of the permutation of initial state, yields various terms: two of them are bilinear in the initial condition $f_0$, and twelve are trilinear in $f_0$. Some of these terms vanish in the limit due to a non-stationary phase argument. Others give rise to truly diverging contributions (negative powers of $\varepsilon$). However, when grouping the terms in the appropriate way, those terms are seen to cancel each other. Last, some terms give the collision operator (4.4). The computation is heavy and hence we address the reader to [5] for the details.

This ends up the analysis of terms up to second order in the potential.

Obviously, as for the Maxwell–Boltzmann case, we could try to re-sum the dominant terms. This would lead to analyzing a true subseries of the complete series expansion expressing $f_j^N(t)$. We do not see any conceptual difficulty, however, this resummation procedure has not been explicitly done in [5].

We mention that a similar analysis, using commutator expansions in the framework of the second quantization formalism, has been performed in [27] (following [26]) in the case of the van Hove limit for lattice systems (that is the same as the weak-coupling limit, yet without rescaling the distances). For more recent formal results in this direction, but in the context of the weak-coupling limit, we also quote [22].

We finally observe that the initial value problem for equation (4.3) is somehow

trivial for Fermions. Indeed we have the a-priori bounds $f \leq \frac{1}{(8\pi)^3}$ making everything easy. For Bosons the situation is much more involved even for the spatially homogeneous case. The statistics favours large values of $f$ and it is not clear whether the equation can explain dynamical condensation. See, for the mathematical side, references [33], [34].

Summarizing, the main scope of this lecture is to show why in the weak-coupling limit, the one-particle distribution function is expected to converge to a solution of Landau equation or Boltzmann equation, for classical and quantum system respectively. From a rigorous view point very little is known.

## 5. Concluding remarks

Other scaling limits yielding different kinetic equations are of course possible. We address the reader to the excellent reference [38] where the various scales and the corresponding kinetic equations are discussed. Here we analyzed in some detail the weak-coupling, however, as mentioned in Section 1, the low-density limit (or the Boltzmann-Grad limit) yields the usual Boltzmann equation for classical systems and this result has been proved for short times. It is natural to investigate what happens, in the same scaling limit, to a quantum system. Here the scaling is

$$t \to \varepsilon t, \quad x \to \varepsilon x, \quad \phi \to \phi, \quad N = \varepsilon^{-2}. \tag{5.1}$$

In other words, the density of obstacles is $\varepsilon$, which is a rarefaction regime, but the potential is unscaled and keeps an $O(1)$ amplitude. Now due to the fact that the density is vanishing, the particles are too rare to make the statistical correlations effective. As a consequence, we expect that the Maxwell–Boltzmann, Bose–Einstein, and Fermi–Dirac situations, all give rise to the same Boltzmann equation along the low-density limit.

As a matter of fact, the expected Boltzmann equation still is a quadratic Boltzmann equation in that case, namely

$$(\partial_t + v \cdot \nabla_x) f(t, x, v) = Q_\ell(f, f)(t, x, v), \tag{5.2}$$

$$Q_\ell(f, f)(t, x, v) \tag{5.3}$$
$$= \int dv_1 \, d\omega \, B_\ell(\omega, v - v_1) \left[ f(t, x, v') f(t, x, v_1') - f(t, x, v) f(t, x, v_1) \right].$$

Here, the index "$\ell$" refers to "low-density".

The factor $B_\ell(\omega, v - v_1)$ is the cross-section. In the low-density limit, collisions take place at a large energy (contrary to the weak-coupling situation), and at a distance of order $\varepsilon$. For this reason, the cross section $B_\ell$ is computed at large energy, and via the quantum rules. In other words, it agrees with the *full* Born series expansion of

quantum scattering, namely

$$B_\ell(\omega, v) = \frac{1}{8\pi^2}|\omega \cdot v|\,|\hat{\phi}(\omega\,(\omega \cdot v))|^2 + \sum_{n\geq 3} B_\ell^{(n)}(\omega, v), \tag{5.4}$$

where each $B_\ell^{(n)}(\omega, v)$ is an explicitly known function, which is $n$-linear in $\phi$ (see [37]). Note that the convergence of the Born series expansion (4.8) is well-known for potentials satisfying a smallness assumption. As it is seen on these formulae, the only difference between the low-density and the weak-coupling regimes (at least for Maxwell–Boltzmann particles) lies in the form of the cross-section. Note also that

$$B_\ell(\omega, v) = B(\omega, v) + O([\phi]^3), \tag{5.5}$$

which reflects the fact that the weak-coupling regime involves only low-energy phenomena.

The analysis of the partial series of the dominant terms (uniform bounds and convergence as for the weak-coupling limit) has been performed in [6] (on the basis of [11] and [12]).

Related problem connected with the ones discussed here are the homogeneization of the distribution function of a single particle in a random distribution of obstacles $\boldsymbol{c} = \{c_1, \ldots, c_N\}$. The basic equations are

$$\dot{x}(t) = v(t), \quad \dot{v}(t) = -\sum_j \nabla\phi(x(t) - c_j) \tag{5.6}$$

for a classical particle and, for a quantum particle

$$i\,\partial_t\psi = -\frac{1}{2}\Delta\psi + \sum_j \phi(x - c_j)\psi. \tag{5.7}$$

We are interested in the behavior of

$$f_\varepsilon(x, v; t) = \mathbb{E}[f_{\boldsymbol{c}}(x, v; t)] \tag{5.8}$$

where $f_{\boldsymbol{c}}(t)$ is the time evolved classical distribution function or the Wigner transform of $\psi$ according to equations (5.6) or (5.7) respectively, under the action of the obstalcle configuration $\boldsymbol{c}$. Finally $\mathbb{E}$ denotes the expectation with respect to the obstacle distribution. For the low-density scaling under a Poisson distribution of obstacles (this is the so called Lorentz model) we obtain, for classical systems, a linear Boltzmann equation (see [23], [39], [3], [17], [10]). It is also known that the system does not homogenize to a jump process given by a linear Boltzmann equation in case of a periodic distribution of obstacles [9]. For the weak-coupling limit we obtain, by a central-limit type of argument, a linear Landau equation as it is shown in [29] and [18].

As regards the corresponding weak-coupling quantum problem, the easiest case is when $\phi$ is a Gaussian process. The kinetic equation is still a linear Boltzmann

equation. The first result, holding for short times, has been obtained in [39] (see also [30]). More recently this result has been extended to arbitrary times [21]. The technique of [21] can been applied to deal with a Poisson distribution of obstacles. Obviously the cross section appearing in the Boltzmann equation is that computed in the Born approximation. Finally in [20] the low-density case has been successfully approached. The result is a linear Boltzmann equation with the full cross-section.

## References

[1] Arseniev, A. A., Buryak, O. E., On a connection between the solution of the Boltzmann equation and the solution of the Landau-Fokker-Planck equation. *Mat. Sb.* **181** (4) (1990), 435–446; English transl. *Math. USSR-Sb.* **69** (2) (1991), 465–478.

[2] Balescu, R., *Equilibrium and Nonequilibrium Statistical Mechanics*. John Wiley & Sons, New York 1975.

[3] Boldrighini, C., Bunimovich, L. A., Sinai, Ya. G., On the Boltzmann equation for the Lorentz gas. *J. Statist. Phys.* **32** (1983), 477–501.

[4] Benedetto, D., Castella, F., Esposito, R., Pulvirenti, M., Some considerations on the derivation of the nonlinear quantum Boltzmann equation. *J. Statist. Phys.* **116** (1–4) (2004), 381–410.

[5] Benedetto, D., Castella, F., Esposito, R., Pulvirenti, M., On the weak-coupling limit for bosons and fermions. *Math. Models Methods Appl. Sci.* **15** (12) (2005), 1–33.

[6] Benedetto, D., Castella, F., Esposito, R., Pulvirenti, M., Some considerations on the derivation of the nonlinear quantum Boltzmann equation II: the low-density regime. *J. Statist. Phys.*, to appear.

[7] Benedetto, D., Castella, F., Esposito, R., Pulvirenti, M., in preparation.

[8] Benedetto, D., Esposito, R., Pulvirenti, M., Asymptotic analysis of quantum scattering under mesoscopic scaling. *Asymptot. Anal.* **40** (2) (2004), 163–187.

[9] Burgain, J., Golse, F., Wennberg, B., On the distribution of free path lenght for the periodic Lorentz gas. *Comm. Math. Phys.* **190** (1998), 491–508.

[10] Caglioti, E., Pulvirenti, M., Ricci, V., Derivation of a linear Boltzmann equation for a lattice gas. *Markov Process. Related Fields* **3** (2000), 265–285

[11] Castella, F., From the von Neumann equation to the quantum Boltzmann equation in a deterministic framework. *J. Statist. Phys.* **104** (1–2) (2001), 387–447.

[12] Castella, F., From the von Neumann equation to the Quantum Boltzmann equation II: identifying the Born series, *J. Statist. Phys.* **106** (5–6) (2002), 1197–1220.

[13] Cercignani, C., *Ludwig Boltzmann. The man who trusted atoms.* Oxford University Press, Oxford 1998.

[14] Cercignani, C., Illner, R., Pulvirenti, M., *The mathematical theory of dilute gases*. Appl. Math. Sci. 106, Springer-Verlag, New York 1994.

[15] Degond, P., Lucquin-Desreux, B., The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case. *Math. Models Methods Appl. Sci.* **2** (2) (1992), 167–182.

[16] DiPerna, R. J., Lions, P.-L., On the Cauchy problem for the Boltzmann equation. *Ann. of Math.* **130** (1989), 321–366.

[17] Desvillettes, L., M. Pulvirenti, M., The linear Boltzmann eqaution for long-range forces: a derivation for nparticle systems *Math. Models Methods Appl. Sci.* **9**, (1999), 1123–1145.

[18] Dürr, D., Goldstain, S., Lebowitz, J. L., Asymptotic motion of a classical particle in a random potential in two dimension: Landau model. *Comm. Math. Phys.* **113** (1987), 209–230.

[19] Esposito, R., Pulvirenti, M., From particles to fluids. In *Handbook of Mathematical Fluid Dynamics* (ed. by S. Friedlander and D. Serre), Vol. 3, North-Holland, Amsterdam 2004, 1–83.

[20] Eng, D., Erdös, L., The linear Boltzmann equation as the low-density limit of a random Schrödinger equation. *Rev. Math. Phys.* **17** (6) (2005), 669–743.

[21] Erdös, L., Yau, H.-T., Linear Boltzmann equation as a weak-coupling limit of a random Schrödinger equation. *Comm. Pure Appl. Math.* **12** (2000), 667–735.

[22] Erdös, L., Salmhofer, M., Yau, H.-T., On the quantum Boltzmann equation. *J. Statist. Phys.* **116** (2004), 367–380.

[23] Gallavotti, G., Rigorous theory of the Boltzmann equation in the Lorentz gas. In *Meccanica Statistica*, reprint, Quaderni CNR 50, 1972, 191–204.

[24] Goudon, T., On Boltzmann equations and Fokker-Planck asymptotics: influence of grazing collisions. *J. Statist. Phys.* **89** (3–4) (1997), 751–776.

[25] Grad, H., On the kinetic Theory of rarefied gases. *Comm. Pure Appl. Math.* **2** (1949), 331–407.

[26] Hugenholtz, M. N., Derivation of the Boltzmann equation for a Fermi gas. *J. Statist. Phys* **32** (1983), 231–254.

[27] Ho, L. N. T., Landau, J., Fermi gas in a lattice in the van Hove limit. *J. Statist. Phys* **87** (1997), 821–845.

[28] Illner, R., Pulvirenti, M., Global Validity of the Boltzmann equation for a two-dimensional rare gas in the vacuum. *Comm. Math. Phys.* **105** (1986), 189–203; Erratum and improved result ibid. **121** (1989), 143–146.

[29] Kesten, H., Papanicolaou, G., A limit theorem for stochastic acceleration *Comm. Math. Phys.* **78**, (1981), 19–31.

[30] Landau, L. J., Observation of quantum particles on a large space-time scale. *J. Statist. Phys* **77** (1994), 259–309.

[31] O. Lanford, O., III, Time evolution of large classical systems. *Dynamical systems, theory and applications* (ed. by E. J. Moser), Lecture Notes in Phys. 38, Springer-Verlag, Berlin 1975, 1–111.

[32] Lifshitz, E. M., Pitaevskii, L. P., *Course of theoretical physics "Landau-Lifshits"*. Vol. 10, Pergamon Internat. Library Sci. Tech. Engrg. Social Stud., Pergamon Press, Oxford-Elmsford, N.Y., 1981.

[33] Lu, Xuguang, The Boltzmann equation for Bose-Einstein particles: velocity concentration and convergence to equilibrium. *J. Statist. Phys.* **119** (2005), 1027–1067.

[34] Lu, Xuguang, On isotropic distributional solutions to the Boltzmann equation for Bose-Einstein particles. *J. Statist. Phys.* **116** (5–6) (2004), 1597–1649.

[35] Nier, F., Une description semi-classique de la diffusion quantique. In *Séminaire sur les Équations aux Dérivées Partielles* ( 1994–1995), Exp. No. VIII, 10 pp., École Polytechnique, Palaiseau 1995.

[36] Nier, F., Asymptotic analysis of a scaled Wigner equation and quantum scattering. *Transport Theory Statist. Phys.* **24** (4–5) (1995), 591–628.

[37] Reed, M., Simon, B., *Methods of modern mathematical physics III. Scattering theory*. Academic Press, New York, London 1979.

[38] Spohn, H., Quantum kinetic equations. In *On Three Levels: Micro-, Meso-, and Macro-Approaches in Physics* (ed. by M. Fannes, C. Maes, A. Verbeure), NATO ASI Series B: Physics 324, Springer-Verlag, Berlin 1994, 1–10.

[39] Spohn, H., The Lorentz flight process converges to a random flight process *Comm. Math. Phys.* **60** (1978), 277–290.

[40] Spohn, H., Derivation of the transport equatin for electrons moving through random impurities *J. Statist. Phys.* **17** (1977), 385–412.

[41] Uehling, E. A., Uhlembeck, G. E., Transport Phenomena in Einstein-Bose and Fermi-Dirac Gases, *Phys. Rev.* **43** (1933), 552–561.

[42] Villani, C., A review of mathematical topics in collisional kinetic theory In *Handbook of Mathematical Fluid Dynamics* (ed. by S. Friedlander and D. Serre), Vol. 1, North-Holland, Amsterdam 2002, 71–307.

[43] Villani, C., On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. *Arch. Rational Mech. Anal.* **143** (3) (1998), 273–307.

[44] Wigner, E., On the quantum correction for the thermodynamical equailibrium. *Phys. Rev.* **40** (1932), 742–759.

Dipartimento di Matematica, Università di Roma, La Sapienza, Roma, Italy
E-mail: pulvirenti@mat.uniroma1.it

# Symmetry of entire solutions for a class of semilinear elliptic equations

Ovidiu Savin

**Abstract.** We discuss a conjecture of De Giorgi concerning the one dimensional symmetry of bounded, monotone in one direction, solutions of semilinear elliptic equations of the form $\Delta u = W'(u)$ in all $\mathbb{R}^n$.

**Mathematics Subject Classification (2000).** 35J70, 35B65.

**Keywords.** Phase transitions models, sliding method, minimal surfaces.

## 1. Introduction

In 1978 De Giorgi [13] made the following conjecture about bounded solutions of a certain semilinear equation:

**Conjecture** (De Giorgi). Let $u \in C^2(\mathbb{R}^n)$ be a solution of

$$\Delta u = u^3 - u, \tag{1}$$

such that

$$|u| \leq 1, \quad u_{x_n} > 0$$

in the whole $\mathbb{R}^n$. Is it true that all the level sets of $u$ are hyperplanes, at least if $n \leq 8$?

The problem originates in the theory of phase transitions and it is closely related to the theory of minimal surfaces. As we explain later, the conjecture is sometimes referred to as "the $\varepsilon$ version of the Bernstein problem for minimal graphs". This relation with the Bernstein problem is the reason why $n \leq 8$ appears in the conjecture.

De Giorgi's conjecture is often considered with the additional natural hypothesis

$$\lim_{x_n \to \pm\infty} u(x', x_n) = \pm 1. \tag{2}$$

Under the much stronger assumption that the limits in (2) are uniform in $x'$, the conjecture is known as the "Gibbons conjecture". This conjecture was first proved for $n \leq 3$ by Ghoussoub and Gui in [18] and then for all dimensions $n$ independently by Barlow, Bass and Gui [4], Berestycki, Hamel and Monneau [6] and Farina [16].

The first positive partial result on the De Giorgi conjecture was established in 1980 by Modica and Mortola [30]. They proved the conjecture in dimension $n = 2$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

under the additional hypothesis that the level sets $\{u = s\}$ are equi-Lipschitz in the $x_2$ direction. Their proof used a Liouville-type theorem for elliptic equations in divergence form, due to Serrin, for the bounded ratio

$$\sigma := \frac{u_{x_1}}{u_{x_2}}.$$

In 1997 Ghoussoub and Gui [18] proved De Giorgi's conjecture for $n = 2$. They used a different Liouville-type theorem for $\sigma$ developed by Berestycki, Caffarelli and Nirenberg in [5] for the study of symmetry properties of positive solutions of semilinear elliptic equations in half spaces. This theorem does not require for $\sigma$ to be bounded, but rather a compatibility condition between the growth of $\sigma$ and the degeneracy of the coefficients of the equation.

Using similar techniques, Ambrosio and Cabre [2] extended these results up to dimension $n = 3$. Also, Ghoussoub and Gui showed in [19] that the conjecture is true for $n = 4$ or $n = 5$ for a special class of solutions that satisfy an anti-symmetry condition.

In 2003 I proved in [33] that the conjecture is true in dimension $n \leq 8$ under the additional hypothesis (2). The proof is nonvariational and uses the sliding method for a special family of radially symmetric functions.

If the level sets of $u$ are assumed to be Lipschitz in the $x_n$ direction, then it was shown by Barlow, Bass and Gui [4] and later in [33] that the solutions are planar in all dimensions.

It is not known whether or not the conjecture is true for all dimensions. Jerison and Monneau [22] showed that the existence of a symmetric minimizer for the energy associated with (1) in $\mathbb{R}^{n-1}$ implies the existence of a counter-example to the conjecture of the De Giorgi in $\mathbb{R}^n$. However, existence of such global minimizer has not been proved.

## 2. Phase transitions

Equations of type (1) arise in variational problems associated with the energy

$$J(u, \Omega) = \int_\Omega \frac{1}{2}|\nabla u|^2 + W(u)\, dx, \quad |u| \leq 1, \tag{3}$$

where $W \in C^2$ is a double well potential with minima at $\pm 1$,

$$W(\pm 1) = W'(\pm 1) = 0, \quad W > 0 \quad \text{on } (-1, 1),$$

$$W''(-1) > 0, \quad W''(1) > 0.$$

We say that $u$ is a local minimizer for $J$ in $\Omega$ if

$$J(u, \Omega) \leq J(u + \varphi, \Omega)$$

for any $\varphi \in C_0^\infty(\Omega)$. Local minimizers of (3) satisfy the Euler–Lagrange equation

$$\Delta u = W'(u). \tag{4}$$

Equation (1) is obtained for the particular choice of the potential

$$W(t) = \frac{1}{4}(1 - t^2)^2.$$

The behavior of minimizers at $\infty$ is given by the properties of blow down solutions

$$u_\varepsilon(x) = u\left(\frac{x}{\varepsilon}\right).$$

These rescalings are local minimizers for the $\varepsilon$ energy functional

$$J_\varepsilon(u_\varepsilon) = \int \frac{\varepsilon}{2}|\nabla u_\varepsilon|^2 + \frac{W(u_\varepsilon)}{\varepsilon}dx.$$

This is a typical energy modeling the phase separation phenomena within the van der Walls–Cahn–Hilliard theory [8]. In this context, $u_\varepsilon$ represents the density of a multi-phase fluid, where the zero points of $W$ correspond to stable fluid phases and the free energy $J_\varepsilon$ depends both on the density potential and the density gradient.

One expects that $u_\varepsilon$ has a transition region of $\mathcal{O}(\varepsilon)$ thickness which approaches a minimal surface as $\varepsilon \to 0$. The intuition behind this comes from the following calculation. By the coarea formula

$$J_\varepsilon(u_\varepsilon, \Omega) \geq \int_\Omega |\nabla u_\varepsilon|\sqrt{2W(u_\varepsilon)}dx = \int_{-1}^{1} \sqrt{2W(s)}\mathcal{H}^{n-1}(\{u_\varepsilon = s\} \cap \Omega)ds.$$

Heuristically, $J_\varepsilon(u_\varepsilon, \Omega)$ is minimized if, in the interior of $\Omega$, the level sets $\{u_\varepsilon = s\}$ are (almost) minimal and

$$|\nabla u_\varepsilon| = \frac{1}{\varepsilon}\sqrt{2W(u_\varepsilon)}. \tag{5}$$

This equation suggests that "the profile" of $u_\varepsilon$ behaves like

$$u_\varepsilon(x) \simeq g\left(\frac{d_{\{u=0\}}(x)}{\varepsilon}\right)$$

where $g$ is the solution of the ordinary differential equation

$$g' = \sqrt{2W(g)}, \quad g(0) = 0,$$

and $d_{\{u=0\}}$ represents the signed distance to the 0 level surface of $u$.

The asymptotic behavior of $u_\varepsilon$ was first studied by Modica and Mortola in [31] and Modica in [25] within the framework of $\Gamma$-convergence. Later, Modica [27], Sternberg [36] and many authors [3], [17], [24], [29], [32], [37] generalized these results for minimizers with volume constraint.

Modica proved in [25] that as $\varepsilon \to 0$, $u_\varepsilon$ has a subsequence

$$u_{\varepsilon_k} \to \chi_E - \chi_{E^c} \quad \text{in } L^1_{\text{loc}} \tag{6}$$

where $E$ is a set with minimal perimeter. Actually, the convergence in (6) is better, as it was shown by Caffarelli and Cordoba in [10]. They proved a uniform density estimate for the level sets of local minimizers $u_\varepsilon$ of $J_\varepsilon$ i.e, if $u_\varepsilon(0) = 0$, then

$$\frac{|\{u_\varepsilon > 0\} \cap B_\delta|}{|B_\delta|} \geq C \tag{7}$$

for $\varepsilon \leq \delta$, $C > 0$ universal. In particular, this implies that in (6) the level sets $\{u_{\varepsilon_k} = \lambda\}$ converge uniformly on compact sets to $\partial E$.

Next we recall some known facts about sets with minimal perimeter.

## 3. Minimal surfaces

The Plateau problem consist in finding a surface of least area (the minimal surface) among those bounded by a given curve. De Giorgi studied this problem by looking at hypersurfaces in $\mathbb{R}^n$ as boundaries of sets. Thus, for a measurable set $E$, he defined the perimeter of $E$ in a domain $\Omega \subset \mathbb{R}^n$ (or the area of $\partial E$ in $\Omega$) as the total variation of $\nabla \chi_E$ in $\Omega$, i.e.

$$P_\Omega(E) = \int_\Omega |\nabla \chi_E| := \sup \left| \int_E \operatorname{div} g \, dx \right|,$$

where the supremum is taken over all vector fields $g \in C_0^1(\Omega)$ with $\|g\|_{L^\infty} \leq 1$.

It is not difficult to show existence to the Plateau problem in this context of minimal boundaries. It is much more difficult to prove that the sets so obtained are actually regular except possibly for a closed singular set.

The main idea to prove "almost everywhere" regularity uses an improvement of flatness theorem due to De Giorgi [14], [21]. Caffarelli and Cordoba gave a different proof in [11] using nonvariational techniques.

**Theorem 3.1** (De Giorgi). *Suppose that $E$ is a set having minimal perimeter in $\{|x'| < 1,\ |x_n| < 1\}$, $0 \in \partial E$ and assume that $\partial E$ is "flat", i.e.*

$$\partial E \subset \{|x_n| < \varepsilon\},$$

*$\varepsilon \leq \varepsilon_0$, $\varepsilon_0$ small universal.*

*Then, possibly in a different system of coordinates, $\partial E$ can be trapped in a flatter cylinder*

$$\{|y'| \leq \eta_2\} \cap \partial E \subset \{|y_n| \leq \varepsilon \eta_1\},$$

*with $0 < \eta_1 < \eta_2$ universal constants.*

This theorem implies that flat minimal surfaces are $C^{1,\alpha}$, and therefore analytic by elliptic regularity theory.

The question of whether or not all points of a minimal surface are regular is closely related to the Bernstein problem. Singular points can exist if and only if there exist nonplanar entire minimal surfaces (or minimal cones). Simons [35] proved that in dimension $n \leq 7$ entire minimal surfaces are planar. Bombieri, De Giorgi and Giusti showed in [7] that the Simons cone

$$\{x \in \mathbb{R}^8 : \ x_1^2 + x_2^2 + x_3^2 + x_4^2 < x_5^2 + x_6^2 + x_7^2 + x_8^2\}$$

is minimal in $\mathbb{R}^8$. Moreover, if the minimal surface is assumed to be a "graph" in some direction, then there are nonplanar minimal graphs only in dimension $n \geq 9$.

Finally we mention that an entire minimal surface that is a "graph" and has at most linear growth at $\infty$ is planar.

## 4. Symmetry of minimizers

It is natural to ask if some properties of minimal surfaces hold also for local minimizers of (3) or solutions of (4). Actually, the conjecture of De Giorgi corresponds to such a question. Results in this direction were obtained by several authors.

Caffarelli and Cordoba proved the uniform density estimate (7) for the level sets of local minimizers $u_\varepsilon$ of $J_\varepsilon$. Modica proved in [28] that solutions of (4) satisfy a monotonicity formula for the energy functional,

$$\frac{J(u, B_R)}{R^{n-1}} \quad \text{increases with } R.$$

I proved in [33] an improvement of flatness theorem for local minimizers of (3) which corresponds to the flatness theorem of De Giorgi for minimal surfaces. It asserts that, if a level set is trapped in a flat cylinder whose height is greater than some given $\theta_0$, then it is trapped in a flatter cylinder in the interior (the flatness depends on $\theta_0$).

**Theorem 4.1** (Savin [33]). *Suppose that $u$ is a local minimizer of* (3) *in the cylinder $\{|x'| < l, \ |x_n| < l\}$, and assume that the* 0 *level set is "flat",*

$$\{u = 0\} \subset \{|x'| < l, \ |x_n| < \theta\},$$

*and contains the point* 0. *Then there exist small constants $0 < \eta_1 < \eta_2 < 1$ depending only on $n$ such that:*

*Given $\theta_0 > 0$ there exists $\varepsilon_1(\theta_0) > 0$ depending on $n$, $W$ and $\theta_0$ such that if*

$$\frac{\theta}{l} \leq \varepsilon_1(\theta_0), \quad \theta_0 \leq \theta$$

*then*

$$\{u = 0\} \cap \{|\pi_\xi x| < \eta_2 l,\ |x \cdot \xi| < \eta_2 l\}$$

*is included in a flatter cylinder*

$$\{|\pi_\xi x| < \eta_2 l,\ |x \cdot \xi| < \eta_1 \theta\}$$

*for some unit vector* $\xi$ ($\pi_\xi$ *denotes the projection along* $\xi$).

The proof uses the fact that at large scales the level sets behave like minimal surfaces and at small scales the equation behaves like Laplace's equation. The ideas are based on a viscosity solution proof of the flatness theorem of De Giorgi (see [34]). As a corollary of the above theorem we obtain that, if the level sets are asymptotically flat at $\infty$ then they are, in fact, hyperplanes.

**Corollary.** *Let* $u$ *be a local minimizer of* $J$ *in* $\mathbb{R}^n$ *with* $u(0) = 0$. *Suppose that there exist sequences of positive numbers* $\theta_k$, $l_k$ *and unit vectors* $\xi_k$ *with*

$$\frac{\theta_k}{l_k} \to 0, \quad l_k \to \infty$$

*such that*

$$\{u = 0\} \cap \{|\pi_{\xi_k} x| < l_k,\ |x \cdot \xi_k| < l_k\} \subset \{|x \cdot \xi_k| < \theta_k\}.$$

*Then the* 0 *level set is a hyperplane.*

*Proof.* Fix $\theta_0 > 0$, and choose $k$ large such that

$$\frac{\theta_k}{l_k} \le \varepsilon \le \varepsilon_1(\theta_0).$$

If $\theta_k \ge \theta_0$ then, by the theorem above, $\{u = 0\}$ is trapped in a flatter cylinder. We apply the theorem repeatedly till the height of the cylinder becomes less than $\theta_0$.

In some system of coordinates we obtain

$$\{u = 0\} \cap \{|y'| < l_k',\ |y_n| < l_k'\} \subset \{|y_n| \le \theta_k'\}$$

with

$$\theta_0 \ge \theta_k' \ge \eta_1 \theta_0, \quad \frac{\theta_k'}{l_k'} \le \frac{\theta_k}{l_k} \le \varepsilon,$$

hence

$$l_k' \ge \frac{\eta_1 \theta_0}{\varepsilon}.$$

We let $\varepsilon \to 0$ and obtain $\{u = 0\}$ is included in an infinite strip of width $\theta_0$. The corollary is proved since $\theta_0$ is arbitrary. □

As a consequence we have the following theorem.

**Theorem 4.2.** *Suppose that $u$ is a local minimizer of $J$ in $\mathbb{R}^n$ and $n \leq 7$. Then the level sets of $u$ are hyperplanes.*

Alberti, Ambrosio and Cabre showed in [1] that monotone solutions of (4) satisfying (2) are in fact local minimizers for the energy (3) (see also the paper of Jerison and Monneau [22]), hence we obtain:

**Theorem 4.3.** *Let $u \in C^2(\mathbb{R}^n)$ be a solution of*

$$\Delta u = W'(u) \tag{8}$$

*such that*

$$|u| \leq 1, \quad \partial_{x_n} u > 0, \quad \lim_{x_n \to \pm\infty} u(x', x_n) = \pm 1. \tag{9}$$

a) *If $n \leq 8$ then the level sets of $u$ are hyperplanes.*

b) *If the 0 level set has at most linear growth at $\infty$ then the level sets of $u$ are hyperplanes.*

The methods developed in [33] are quite general and can be applied for other types of nonlinear, possibly degenerate elliptic equations. Recently Valdinoci, Sciunzi and Savin [40] proved the theorems above for the energy

$$J_p(u, \Omega) := \int_\Omega \frac{1}{p} |\nabla u|^p + W(u) dx,$$

and the corresponding $p$-Laplace equation

$$\Delta_p u = W'(u), \quad \Delta_p u := \operatorname{div}(|\nabla u|^{p-2} \nabla u).$$

Similar results can be obtained for solutions of nonlinear reaction–diffusion equations of the type

$$F(D^2 u) = f(u), \quad u \in C^2(\mathbb{R}^n), \; u_{x_n} > 0, \; \lim_{x_n \to \pm\infty} u(x', x_n) = \pm 1, \tag{10}$$

where $F$ is uniformly elliptic, and $F$, $f$ are such that there exists a one dimensional solution $g$ which solves the equation in all directions, i.e.,

$$F(D^2 g(x \cdot \nu) = f(g(x \cdot \nu)), \quad \text{for all } \nu \in \mathbb{R}^n, |\nu| = 1.$$

If the rescaled level sets $\varepsilon_k \{u = 0\}$ converge uniformly on compact sets as $\varepsilon_k \to 0$, then the limiting surface satisfies a uniformly elliptic equation (depending on $F$, $f$) for its second fundamental form (instead of the minimal surface equation). Following ideas from [33] one can prove a Liouville property for level sets of solutions of (10). More precisely, if $\{u = 0\}$ stays above a hyperplane $x \cdot \nu = \text{const.}$, then $u$ depends only on one variable i.e,

$$u(x) = g(x \cdot \nu).$$

Moreover, if $F$ is smooth, then the level sets satisfy an improvement of the flatness theorem. This implies that if the 0 level set is asymptotically flat at $\infty$, then $u$ depends only on one variable. In particular, if the 0 level set is a Lipschitz graph in the $x_n$ direction, then the same conclusion holds.

## References


[1] Alberti, G., Ambrosio, L., Cabre, X., On a long-standing conjecture of E. De Giorgi: symmetry in 3D for general nonlinearities and a local minimality property. *Acta Appl. Math.* **65** (1–3) (2001), 9–33.

[2] Ambrosio, L., Cabre, X., Entire solutions of semilinear elliptic equations in $\mathbb{R}^3$ and a conjecture of De Giorgi. *J. American Math. Soc.* **13** (2000), 725–739.

[3] Baldo, S., Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **7** (2) (1990), 67–90.

[4] Barlow, M., Bass, R., Gui, C., The Liouville property and a conjecture of De Giorgi. *Comm. Pure Appl. Math.* **53** (8) (2000), 1007–1038.

[5] Berestycki, H., Caffarelli, L., Nirenberg, L., Further qualitative properties for elliptic equations in unbounded domains. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **25** (1–2) (1997), 69–94.

[6] Berestycki, H., Hamel, F., Monneau,, R., One-dimensional symmetry of bounded entire solutions of some elliptic equations. *Duke Math. J.* **103** (3) (2000), 375–396.

[7] Bombieri, E., De Giorgi, E., Giusti, E., Minimal cones and the Bernstein problem. *Invent. Math.* **7** (1969), 243–268.

[8] Cahn, J., Hilliard, J., Free energy of a nonuniform system I. Interfacial free energy. *J. Chem. Phys* **28** (1958), 258–267.

[9] Caffarelli, L., Cabre, X., *Fully Nonlinear Elliptic Equations*. Amer. Math. Soc. Colloq. Publ. 43, Amer. Math. Soc., Providence, RI, 1995.

[10] Caffarelli, L., Cordoba, A., Uniform convergence of a singular perturbation problem. *Comm. Pure Appl. Math.* **48** (1) (1995), 1–12.

[11] Caffarelli, L., Cordoba, A., An elementary regularity theory of minimal surfaces. *Differential Integral Equations* **6** (1) (1993), 1–13.

[12] Caffarelli, L., Garofalo, N., Segla, F., A gradient bound for entire solutions of quasi-linear equations and its consequences. *Comm. Pure Appl. Math.* **47** (11) (1994), 1457–1473.

[13] De Giorgi, E., Convergence problems for functional and operators. In *Proc. Intern. Meeting on Recent Methods in Nonlinear Analysis* (Rome, 1978), Pitagora, Bologna 1979, 131–188.

[14] De Giorgi, E., *Frontiere orientate di misura minima*. Sem. Mat. Scuola Norm. Sup. Pisa 1960, Editrice Tecnico Scientifica, Pisa 1961.

[15] Farina, A., Some remarks on a conjecture of De Giorgi. *Calc. Var. Partial Differential Equations* **8** (1999), 129–154.

[16] Farina, A., Symmetry for solutions of semilinear elliptic equations in $\mathbb{R}^n$ and related conjectures. *Ricerche Mat.* **48** (1999), 129–154.

[17] Fonseca, I., Tartar, L., The gradient theory of phase transitions for systems with two potential wells. *Proc. Roy. Soc. Edinburgh Sect. A* **111** (1–2) (1989), 89–102.

[18] Ghoussoub, N., Gui, C., On a conjecture of De Giorgi and some related problems. *Math. Ann.* **311** (1998), 481–491.

[19] Ghoussoub, N., Gui, C., On De Giorgi's conjecture in dimensions 4 and 5. *Ann. of Math.* (2) **157** (1) (2003), 313–334.

[20] Gilbarg, D., Trudinger, N., *Elliptic Partial Differential Equations of second order*. Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, Heidelberg, New York 1983.

[21] Giusti, E., Minimal Surfaces and functions of bounded variation, Birkhäuser Verlag, Basel, Boston 1984.

[22] Jerison, D., Monneau, R., The existence of a symmetric global minimizer on $\mathbb{R}^{n-1}$ implies the existence of a counter-example to a conjecture of De Giorgi in $\mathbb{R}^n$. *C. R. Acad. Sci. Paris Sér. I Math* **333** (5) (2001), 427–431.

[23] Kawohl, B., Symmetry or not? *Math. Intelligencer* **20** (1998), 16–22.

[24] Kohn, R., Sternberg, P., Local minimisers and singular perturbations. *Proc. Roy. Soc. Edinburgh Sect. A* **111** (1–2) (1989), 69–84.

[25] Modica, L., $\Gamma$-convergence to minimal surfaces problem and global solutions of $\Delta u = 2(u^3 - u)$. In *Proc. Intern. Meeting on Recent Methods in Nonlinear Analysis* (Rome, 1978), Pitagora, Bologna 1979, 223–244.

[26] Modica, L., A gradient bound and a Liouville theorem for nonlinear Poisson equations. *Comm. Pure Appl. Math.* **38** (5) (1985), 679–684.

[27] Modica, L., The gradient theory of phase transitions and the minimal interface criterion. *Arch. Rational Mech. Anal.* **98** (2) (1987), 123–142.

[28] Modica, L., Monotonicity of the energy for entire solutions of semilinear elliptic equations. In *Partial differential equations and the calculus of variations*, Vol. II, Progr. Nonlinear Differential Equations Appl. 2, Birkhäuser, Boston, MA, 1989, 843–850.

[29] Modica, L., Gradient theory of phase transitions with boundary contact energy. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **4** (5) (1987), 487–512.

[30] Modica, L., Mortola, S., Some entire solutions in the plane of nonlinear Poisson equations. *Boll. Un. Mat. Ital. B (5)* **17** (2) (1980), 614–622.

[31] Modica, L., Mortola, S., Un esempio di $\Gamma^-$-convergenza. *Boll. Un. Mat. Ital. B* (5) **14** (1) (1977), 285–299.

[32] Owen, N., Nonconvex variational problems with general singular perturbations. *Trans. Amer. Math. Soc.* **310** (1) (1988), 393–404.

[33] Savin, O., Phase transitions: Regularity of flat level sets. PhD Thesis, UT Austin, 2003.

[34] Savin, O., Small perturbation solutions for elliptic equations. *Comm. Partial Differential Equations*, to appear.

[35] Simons, J., Minimal varieties in riemannian manifolds. *Ann. of Math.* (2) **88** (1968), 62–105.

[36] Sternberg, P., The effect of a singular perturbation on nonconvex variational problems. *Arch. Rational Mech. Anal.* **101** (3) (1988), 209–260.

[37] Sternberg, P., Vector-valued local minimizers of nonconvex variational problems. *Rocky Mountain J. Math.* **21** (2) (1991), 799–807.

[38] Tonegawa, Y., Phase field model with a variable chemical potential. *Proc. Roy. Soc. Edinburgh Sect. A* **132** (4) (2002), 993–1019.

[39] Tonegawa, Y., Domain dependent monotonicity formula for a singular perturbation problem. *Indiana Univ. Math. J.* **52** (1) (2003), 69–83.

[40] Valdinoci, E., Sciunzi, B., Savin, O., Flat level set regularity of $p$-Laplace phase transitions. Mem. Amer. Math. Soc., to appear.

Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720, U.S.A.

E-mail: osavin@math.berkeley.edu

# Vortices in the Ginzburg–Landau model of superconductivity

Sylvia Serfaty

**Abstract.** We review some mathematical results on the Ginzburg–Landau model with and without magnetic field. The Ginzburg–Landau energy is the standard model for superconductivity, able to predict the existence of vortices (which are quantized, topological defects) in certain regimes of the applied magnetic field. We focus particularly on deriving limiting (or reduced) energies for the Ginzburg–Landau energy functional, depending on the various parameter regimes, in the spirit of $\Gamma$-convergence. These passages to the limit allow to perform a sort of dimension-reduction and to deduce a rather complete characterization of the behavior of vortices for energy-minimizers, in agreement with the physics results. We also describe the behavior of energy critical points, the stability of the solutions, the motion of vortices for solutions of the gradient-flow of the Ginzburg–Landau energy, and show how they are also governed by those of the limiting energies.

**Mathematics Subject Classification (2000).** Primary 00A05; Secondary 00B10.

**Keywords.** Ginzburg–Landau equations, variational methods, $\Gamma$-convergence, superconductivity, vortices.

## 1. Introduction

**1.1. Presentation of the Ginzburg–Landau model.** We are interested in describing mathematical results on the two-dimensional Ginzburg–Landau model. This is a model of great importance and recognition in physics (with several Nobel prizes awarded for it: Landau, Ginzburg, Abrikosov). It was introduced by Ginzburg and Landau (see [15]) in the 1950s as a phenomenological model to describe superconductivity. Superconductivity was itself discovered in 1911 by Kammerling Ohnes. It consists in the complete loss of resistivity of certain metals and alloys at very low temperatures. The two most striking consequences of it are the possibility of permanent *superconducting currents* and the particular behavior that, when the material is submitted to an external magnetic field, that field gets expelled from it. Aside from explaining these phenomena, and through the very influential work of Abrikosov [1], the Ginzburg–Landau model allowed to predict the possibility of a *mixed state* in type-II superconductors where triangular vortex lattices appear. These vortices – a vortex can be described in a few words as a quantized amount of vorticity of the superconducting current localized near a point – have since been the objects of many observations and experiments.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The Ginzburg–Landau theory has also been justified as a limit of the Bardeen–Cooper–Schrieffer (BCS) quantum theory, which explains superconductivity by the existence of "Cooper pairs" of superconducting electrons.

In addition to its importance in the modelling of superconductivity, the Ginzburg–Landau model turns out to be the simplest case of a gauge theory, and vortices to be the simplest case of topological solitons (for these aspects see [34], [21] and the references therein); moreover, it is mathematically extremely close to the Gross–Pitaevskii model for superfluidity and models for rotating Bose–Einstein condensates in which quantized vortices are also essential objects, to which the Ginzburg–Landau techniques have been successfully exported.

The 2D Ginzburg–Landau model leads (after various suitable rescalings) to describing the state of the superconducting sample submitted to the external field $h_{\text{ex}}$, below the critical temperature, through the energy functional

$$G_\varepsilon(u, A) = \frac{1}{2}\int_\Omega |\nabla_A u|^2 + |\text{curl } A - h_{ex}|^2 + \frac{1}{2\varepsilon^2}\left(1 - |u|^2\right)^2. \qquad (1)$$

In this expression, $\Omega$ is a two-dimensional open subset of $\mathbb{R}^2$, which in our study is always assumed for simplicity to be smooth, bounded and simply connected. One can imagine it represents the section of an infinitely long cylinder, or a thin film.

The first unknown $u$ is a *complex*-valued function, called "order parameter" in physics, where it is generally denoted $\psi$. It is a condensed wave function, indicating the local state of the material or the phase in the Landau theory approach of phase transitions: $|u|^2$ is the density of "Cooper pairs" of superconducting electrons (responsible for superconductivity in the BCS approach). With our normalization $|u| \leq 1$ and where $|u| \simeq 1$ the material is in the superconducting phase, while where $|u| = 0$, it is in the normal phase (i.e. behaves like a normal conductor), the two phases being able to coexist in the sample.

The second unknown is $A$, the electromagnetic vector-potential of the magnetic field, a function from $\Omega$ to $\mathbb{R}^2$. The magnetic field in the sample is deduced by $h = \text{curl } A = \partial_1 A_2 - \partial_2 A_1$, it is thus a real-valued function in $\Omega$. The notation $\nabla_A$ denotes the covariant gradient $\nabla - iA$; $\nabla_A u$ is thus a vector with complex components.

The *superconducting current* is a real vector given by $\langle iu, \nabla_A u\rangle$ where $\langle .\,, .\rangle$ denotes the scalar-product in $\mathbb{C}$ identified with $\mathbb{R}^2$, it may also be written as $\frac{i}{2}\left(u\overline{\nabla_A u} - \bar{u}\nabla_A u\right)$, where the bar denotes complex conjugation.

The parameter $h_{\text{ex}} > 0$ represents the intensity of the applied field (assumed to be perpendicular to the plane of $\Omega$). Finally, the parameter $\varepsilon$ is the inverse of the "Ginzburg–Landau parameter" usually denoted $\kappa$, a non-dimensional parameter depending on the material only. We will be interested in the regime of small $\varepsilon$ or $\kappa \to +\infty$, corresponding to high-$\kappa$ superconductors (also called the London limit). In this limit, the characteristic size of the vortices, which is $\varepsilon$, tends to 0 and vortices become point-like.

The stationary states of the system are the critical points of $G_\varepsilon$, or the solutions of the Ginzburg–Landau equations (Euler–Lagrange equations associated to $G_\varepsilon$):

$$\text{(GL)}\quad \begin{cases} -(\nabla_A)^2 u = \dfrac{1}{\varepsilon^2} u(1-|u|^2) & \text{in } \Omega \\ -\nabla^\perp h = \langle iu, \nabla_A u\rangle & \text{in } \Omega \\ h = h_{\text{ex}} & \text{on } \partial\Omega \\ \nu \cdot \nabla_A u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\nabla^\perp$ denotes the operator $(-\partial_2, \partial_1)$, and $\nu$ the outer unit normal to $\partial\Omega$.

The Ginzburg–Landau equations and functional are invariant under $\mathbb{U}(1)$-gauge transformations (it is an Abelian gauge-theory):

$$\begin{cases} u \mapsto u e^{i\Phi}, \\ A \mapsto A + \nabla\Phi. \end{cases} \tag{2}$$

The physically relevant quantities are those that are gauge-invariant, such as the energy $G_\varepsilon$, $|u|$, $h$, etc.

For more on the model and on the physics, we refer to the physics literature, in particular [53], [13]. For more reference on the results we present here, we refer to our monograph with E. Sandier [47].

We will also mention results on the simplified Ginzburg–Landau model, without magnetic field. It consists in taking $A = 0$ and $h_{\text{ex}} = 0$, then the energy reduces to

$$E_\varepsilon(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 + \frac{(1-|u|^2)^2}{2\varepsilon^2} \tag{3}$$

with still $u\colon \Omega \to \mathbb{C}$. Critical points of this energy are solutions of

$$-\Delta u = \frac{u}{\varepsilon^2}(1-|u|^2). \tag{4}$$

It is a complex-valued version of the Allen–Cahn model for phase-transitions. The first main study of this functional was done by Bethuel–Brezis–Hélein in the book [6], where they replace the effect of the applied field $h_{\text{ex}}$ by a fixed Dirichlet boundary condition. Since then, this model has been extensively studied.

**1.2. Vortices.** A vortex is an object centered at an isolated zero of $u$, around which the phase of $u$ has a nonzero winding number, called the *degree of the vortex*. When $\varepsilon$ is small, it is clear from (1) that $|u|$ prefers to be close to 1, and a scaling argument hints that $|u|$ is different from 1 in regions of characteristic size $\varepsilon$. A typical vortex centered at a point $x_0$ "looks like" $u = \rho e^{i\varphi}$ with $\rho(x_0) = 0$ and $\rho = f(\frac{|x-x_0|}{\varepsilon})$ where $f(0) = 0$ and $f$ tends to 1 as $r \to +\infty$, i.e. its characteristic core size is $\varepsilon$, and

$$\frac{1}{2\pi}\int_{\partial B(x_0, R\varepsilon)} \frac{\partial\varphi}{\partial\tau} = d \in \mathbb{Z}$$

is an integer, called the *degree of the vortex*. For example $\varphi = d\theta$ where $\theta$ is the polar angle centered at $x_0$ yields a vortex of degree $d$. We have the important relation

$$\operatorname{curl} \nabla\varphi = 2\pi \sum_i d_i \delta_{a_i} \tag{5}$$

where the $a_i$'s are the centers of the vortices, the $d_i$'s their degrees, and $\delta$ the Dirac mass.

In the limit $\varepsilon \to 0$ vortices become *point-like* or more generally, in any dimension, *codimension* 2 singularities (see [30], [7]) – to be compared with the case of real-valued phase-transition models (Allen–Cahn), where the order parameter $u$ is real-valued, leading to codimension 1 singular sets in the limit.

**1.3. Critical fields.** When an external magnetic field is applied to a superconductor, several responses can be observed depending on the intensity of the field $h_{\text{ex}}$.

There are three main critical values of $h_{\text{ex}}$ or *critical fields* $H_{c_1}$, $H_{c_2}$, and $H_{c_3}$, for which phase-transitions occur. Below the first critical field, which is of order $O(|\log \varepsilon|)$ (as first established by Abrikosov), the superconductor is everywhere in its superconducting phase $|u| \sim 1$ and the magnetic field does not penetrate (this is called the Meissner effect or Meissner state). At $H_{c_1}$, the first vortice(s) appear. Between $H_{c_1}$ and $H_{c_2}$, the superconducting and normal phases (in the form of vortices) coexist in the sample, and the magnetic field penetrates through the vortices. This is called the *mixed state*. The higher $h_{\text{ex}} > H_{c_1}$, the more vortices there are. Since they repel each other, they tend to arrange in these triangular Abrikosov lattices in order to minimize their repulsion. Reaching $H_{c_2} \sim \frac{1}{\varepsilon^2}$, the vortices are so densely packed that they overlap each other, and at $H_{c_2}$ a second phase transition occurs, after which $|u| \sim 0$ inside the sample, i.e. all superconductivity in the bulk of the sample is lost.

In the interval $[H_{c_2}, H_{c_3}]$ however, superconductivity persists near the boundary, this is called *surface superconductivity*. Above $H_{c_3} = O\left(\frac{1}{\varepsilon^2}\right)$ (defined in decreasing fields), the sample is completely in the normal phase $u \equiv 0$, the magnetic field completely penetrates, and decreasing the field below $H_{c_3}$, surface superconductivity is observed.

**1.4. Questions, results and methods.** The main question is to understand mathematically the behavior above, and in particular:

– To understand the vortices and their repartition, interaction.

– To understand the influence of the boundary conditions and/or of the applied field.

– To find the asymptotic values of the critical fields (as $\varepsilon \to 0$).

– To prove compactness results and derive *limiting energies/reduced problems*, thus following the strategy of $\Gamma$-convergence. This enables to understand the behavior of global minimizers (or energy minimizers) and their vortices. In order to achieve

this, one needs to find *lower bounds* for the energy, together with matching upper bounds.

– To understand and find local minimizers. This is done through a special "local minimization in energy sectors" method.

– To understand the behavior of critical points of the energy (i.e. solutions which are not necessarily stable), that is to pass to the limit $\varepsilon \to 0$ in the Ginzburg–Landau equations (GL). The method used here is to pass to the limit in the "stress-energy tensor".

– To derive the limiting motion law of vortices, and to understand its link with the reduced energies mentioned above.

## 2. Mathematical tools

Various methods were introduced to describe vortices, since [6]. A crucial difference between the analysis for (1) and the one for (3) is that for the case with magnetic field (1) we really need to be able to handle numbers of vortices which are *unbounded* as $\varepsilon \to 0$. We designed tools able to capture vortices for arbitrary maps $u$ (not necessarily solutions), and to treat possibly unbounded numbers.

Let us describe the two main technical tools which we use throughout: the *"vortex ball construction"*, yielding the lower bounds on the energy, and the *vorticity measures*, which serve to describe vortex-densities instead of individual vortices.

**2.1. The vortex-ball construction.** This serves to obtain lower bounds for bounded or unbounded numbers of vortices. The idea is that, whatever the map $u$, for topological reasons, a vortex of degree $d$ confined in a ball of radius $r$ should cost at least an energy $\pi d^2 \log \frac{r}{\varepsilon}$. Then, since there may be a large number of these vortices, one must find a way to add up those estimates. It is done following the ball-growth method of Sandier [38] and Jerrard [22] (which consists in growing annuli of same conformal type and merging them appropriately). The best result to date is the following:

**Theorem 2.1** (see [47]). *Let $(u, A)$ be a configuration such that $E_\varepsilon(|u|) \leq C\varepsilon^{\alpha-1}$ with $\alpha > 0$, then for any $r \in (\varepsilon^{\frac{\alpha}{2}}, 1)$, and $\varepsilon$ small enough, there exists a finite collection of disjoint closed balls $\{B_i\}_i$ of centers $a_i$, of sum of the radii $r$, covering $\{|u| \leq 1 - \varepsilon^{\frac{\alpha}{4}}\} \cap \{x \in \Omega, \operatorname{dist}(x, \partial\Omega) \geq \varepsilon\}$, and such that*

$$\frac{1}{2}\int_{\bigcup_i B_i} |\nabla_A u|^2 + |\operatorname{curl} A|^2 + \frac{(1-|u|^2)^2}{2\varepsilon^2} \geq \pi D\left(\log\frac{r}{D\varepsilon} - C\right) \tag{6}$$

*where $D = \sum_i |d_i|$ and $d_i = \deg(u, \partial B_i)$.*

This lower bound is very general, it does not require any hypothesis on $(u, A)$ other than a reasonable (but quite large) upper bound on its energy, and it is in fact sharp (examples where it is can be constructed).

**2.2. The vorticity measures.** Recall that a complex-valued map $u$ can be written in polar coordinates $u = \rho e^{i\varphi}$ with a phase $\varphi$ which can be multi-valued. Given a configuration $(u, A)$, we define its *vorticity* by

$$\mu(u, A) = \operatorname{curl} \langle iu, \nabla_A u \rangle + \operatorname{curl} A. \tag{7}$$

Formally

$$\langle iu, \nabla u \rangle = \rho^2 \nabla \varphi \simeq \nabla \varphi,$$

considering that $\rho = |u| \simeq 1$. Taking the curl of this expression and using (5), one would get the approximate (formal) relation

$$\mu(u, A) \simeq 2\pi \sum_i d_i \delta_{a_i} \tag{8}$$

where $a_i$'s are the vortices of $u$ and $d_i$'s their degrees. Thus we see why the quantity $\mu$ corresponds to a vorticity-measure of the map $u$ (just like the vorticity for fluids). The following theorem gives a rigorous content to (8).

**Theorem 2.2** (see [24] and [47]). *The $(a_i, d_i)'$s being given by the previous theorem, we have*

$$\left\| \mu(u, A) - 2\pi \sum_i d_i \delta_{a_i} \right\|_{(C_0^{0,\gamma}(\Omega))^*} \leq C r^\gamma G_\varepsilon^0(u, A) \quad \textit{for all } 0 < \gamma < 1,$$

*where $G_\varepsilon^0$ is the energy when $h_{\text{ex}} = 0$.*

The previous theorem allowed to give a control on the mass of $2\pi \sum_i d_i \delta_{a_i}$ as measures. This one ensures that if $r$ is taken small enough, $\mu(u, A)$ and $2\pi \sum_i d_i \delta_{a_i}$ are close in a weak norm. Combining the two yields a compactness result on the vorticity $\mu(u, A)$, if rescaled by dividing by the number of vortices. It also ensures the limiting vorticity is a bounded Radon measure. This is the limiting "vortex-density" we are looking to characterize in various situations.

**Remark 2.3.** When the Ginzburg–Landau equations (GL) are satisfied, taking the curl of the second relation, we find that the vorticity and the induced field are linked by the relation

$$\begin{cases} -\Delta h + h = \mu(u, A) & \text{in } \Omega, \\ h = h_{\text{ex}} & \text{on } \partial\Omega. \end{cases} \tag{9}$$

Thus the knowledge of the vorticity is equivalent to that of the induced field $h$.

## 3. Global minimization ($\Gamma$-convergence type) results

### 3.1. Results for $E_\varepsilon$

**3.1.1. In two dimensions.** For the two-dimensional simplified model (3), the main result of [6] can be written in the following form.

**Theorem 3.1** (Bethuel–Brezis–Hélein [6]). *Let $\Omega$ be a strictly starshaped simply connected domain of $\mathbb{R}^2$ and $g\colon \partial\Omega \to \mathbb{S}^1$ a smooth map of degree $d > 0$. If $u_\varepsilon$ minimizes $E_\varepsilon$ among maps with values $g$ on $\partial\Omega$, then, as $\varepsilon \to 0$, up to extraction of a subsequence, there exist $d$ distinct points $a_1, \dots, a_d \in \Omega$ such that $u_\varepsilon \to u_*$ in $C^k_{\mathrm{loc}}\big(\Omega \setminus \bigcup_i \{a_i\}\big)$ where*

1. *$u_*$ is a harmonic map from $\Omega\setminus\{a_1, \dots, a_d\}$ to $\mathbb{S}^1$ with $u_* = g$ on $\partial\Omega$ and with degree $d_i = 1$ around each $a_i$,*
2. *$(a_1, \dots, a_d)$ is a minimizer of the renormalized energy $W$ with $d_i = 1$,*
3. *$E_\varepsilon(u_\varepsilon) \geq \pi d|\log\varepsilon| + W(a_1, \dots, a_d) + o(1)$.*

Here $W$ denotes a function of the points $a_i \in \Omega$ (depending also on the degrees), called *"renormalized energy"* and which has a form

$$W(x_1, \dots, x_n) = -\pi \sum_{i\neq j} d_i d_j \log|x_i - x_j| + \text{terms of interaction with the boundary.}$$

$W$ corresponds to the finite part of the energy left when subtracting the "infinite" self-interaction cost of the vortices $\pi d|\log\varepsilon|$, i.e. to the interaction between the vortices (vortices of same sign repel, of opposite sign attract).

This result can be phrased as a $\Gamma$-convergence result (in the sense of DeGiorgi):

**Proposition 3.2** ($\Gamma$-convergence of $E_\varepsilon$).

1. *For any family $\{u_\varepsilon\}_\varepsilon$ such that $E_\varepsilon(u_\varepsilon) \leq C|\log\varepsilon|$ and $u_\varepsilon = g$ on $\partial\Omega$; up to extraction, there exists a finite family $(a_i, d_i)$ of $n$ points + degrees such that $\sum_{i=1}^n d_i = d$ and*

$$\operatorname{curl}\langle i u_\varepsilon, \nabla u_\varepsilon\rangle \rightharpoonup 2\pi \sum_{i=1}^n d_i \delta_{a_i},$$

$$E_\varepsilon(u_\varepsilon) \geq \pi \sum_{i=1}^n |d_i||\log\varepsilon| + W(a_1, \dots, a_n) + o(1) \quad \text{as } \varepsilon \to 0.$$

2. *For all distinct $a_i$'s and $d_i = \pm 1$, there exists $u_\varepsilon$ such that*

$$E_\varepsilon(u_\varepsilon) \leq \pi n|\log\varepsilon| + W(a_1, \dots, a_n) + o(1).$$

Fixing the degrees $d_i = \pm 1$ and the number of vortices $n$, this result states exactly that $E_\varepsilon - \pi n |\log \varepsilon|$ $\Gamma$-converges to $W$. This reduces the dimension of the problem, by reducing the minimization of $E_\varepsilon$ to the simple one of the limiting energy $W$, which is a function on a finite dimensional set.

**3.1.2. In higher dimensions.** Three-dimensional as well as higher-dimensional versions of Theorem 3.1 have been given, in particular by Lin–Rivière [30], Sandier [39], Bethuel–Brezis–Orlandi [7]. Jerrard and Soner gave a $\Gamma$-convergence formulation (i.e. analogous to Proposition 3.2), later improved by Alberti–Baldo–Orlandi [2]. Here $E_\varepsilon$ refers to the $n$-dimensional version of the energy (3). When $n = 3$ the vortex-set (or zero-set of $u$) is a set of lines, in higher dimensions it is a set of codimension 2, and the vorticity is best described in the language of currents.

**Theorem 3.3** (Jerrard–Soner [24]). *Let $\{u_\varepsilon\}_\varepsilon$ be a family such that $E_\varepsilon(u_\varepsilon) \leq C|\log \varepsilon|$; up to extraction, there exists an integer-multiplicity rectifiable $(n-2)$-dimensional current $J$ such that*

$$\mu_\varepsilon(u_\varepsilon) := *d\langle i u_\varepsilon, d u_\varepsilon\rangle \rightharpoonup 2\pi J \quad \text{in } (C_0^{0,\gamma}(\Omega))^*$$

*for all $\gamma < 1$ (in the language of differential forms) and*

$$\varliminf_{\varepsilon\to 0} \frac{E_\varepsilon(u_\varepsilon)}{|\log\varepsilon|} \geq \pi \|J\|(\Omega),$$

*where $\|J\|(\Omega)$ is the total mass of the current.*

The total mass of the current corresponds in dimension 3 to the total length of the vortex-lines. The result of [30], [39] essentially states that minimizers of $E_\varepsilon$ have a vorticity $\mu_\varepsilon$ which converges to minimizers of the length $\|J\|(\Omega)$, thus have vortices which converge to straight lines or minimal connections (or codimension 2 minimal currents in higher dimension). The result of [7] generalizes it to critical points and proves that critical points of $E_\varepsilon$ have vortices which converges to stationary varifolds.

Observe that the situation is quite different from the dimension 2, because the main order $|\log\varepsilon|$ of the energy already gives a nontrivial limiting problem: the mass of the limiting object $J$; in contrast with the 2D problem which only leads to minimizing the number of points (one needs to go to the next order in the energy to get an interesting problem: the minimization of $W$).

### 3.2. Global minimization results for $G_\varepsilon$

**3.2.1. Close to $H_{c_1}$.** Let us introduce $h_0$ the solution of

$$\begin{cases} -\Delta h_0 + h_0 = 0 & \text{in } \Omega, \\ h_0 = 1 & \text{on } \partial\Omega, \end{cases} \tag{10}$$

and

$$C(\Omega) = (2 \max |h_0 - 1|)^{-1}. \tag{11}$$

We also introduce the set $\Lambda = \{x \in \Omega / h_0(x) = \min h_0\}$ and we will assume here for simplicity that it is reduced to only one point called $p$, and denote $Q(x) = \langle D^2 h_0(p)x, x\rangle$, assumed to be definite positive. With these notations, a first essential result is the asymptotic formula for $H_{c_1}$ (confirming physical predictions that $H_{c_1} = O(|\log \varepsilon|)$):

$$H_{c_1} = C(\Omega)|\log \varepsilon| + O(1). \tag{12}$$

**Theorem 3.4** ([48], [47]). *Assume* $h_{\text{ex}} \leq H_{c_1} + O(\log\ |\log \varepsilon|)$, *then for* $h_{\text{ex}} \in (H_n, H_{n+1})$ *where* $H_n$ *has the expansion*

$$H_n = C(\Omega)\left(|\log \varepsilon| + (n-1)\log \frac{|\log \varepsilon|}{n}\right) + \textit{lower order terms,}$$

*global minimizers of* $G_\varepsilon$ *have exactly n vortices of degree* 1, $a_i^\varepsilon \to p$ *as* $\varepsilon \to 0$ *and the* $\tilde{a_i^\varepsilon} = \sqrt{\frac{h_{\text{ex}}}{n}}(a_i^\varepsilon - p)$ *converge as* $\varepsilon \to 0$ *to a minimizer of*

$$w_n(x_1, \ldots, x_n) = -\pi \sum_{i \neq j} \log |x_i - x_j| + \pi n \sum_{i=1}^{n} Q(x_i). \tag{13}$$

Through this theorem we see that the behavior is as expected: below $H_{c_1} = H_1$ there are no vortices in energy minimizers (in addition it was proved in [49] that the minimizer is unique), then at $H_{c_1}$ the first vortex becomes favorable, close to the point $p$. Then, there is a sequence of additional critical fields $H_2, H_3, \ldots$ separated by increments of $\log\ |\log \varepsilon|$, for which a second, third, etc, vortex becomes favorable. Each time the optimal vortices are located close to $p$ as $\varepsilon \to 0$, and after blowing-up at the scale $\sqrt{\frac{h_{\text{ex}}}{n}}$ around $p$, they converge to configurations which minimize $w_n$ in $\mathbb{R}^2$. Now, $w_n$, which appears as a limiting energy (after that rescaling) contains a repulsion term like $W$, and a confinement term due to the applied field. It is a standard two-dimensional interaction, however rigorous results on its minimization are hard to obtain as soon as $n \geq 3$. When $Q$ has rotational symmetry, numerical minimization (see Gueron–Shafrir [17]) yields very regular shapes (regular polygons for $n \leq 6$, regular stars) which look very much like the birth of a triangular lattice as $n$ becomes large. All these results are in very good agreement with experimental observations.

**3.2.2. Global minimizers in the intermediate regime.** In the next higher regime of applied field, the result is the following:

**Theorem 3.5** ([47]). *Assume* $h_{\text{ex}}$ *satisfies, as* $\varepsilon \to 0$,

$$\log\ |\log \varepsilon| \ll h_{\text{ex}} - H_{c_1} \ll |\log \varepsilon|.$$

*Then there exists* $1 \ll n_\varepsilon \ll h_{\text{ex}}$ *such that*

$$h_{\text{ex}} \sim C(\Omega)\left(|\log\varepsilon| + n_\varepsilon \log\frac{|\log\varepsilon|}{n_\varepsilon}\right)$$

*and if* $(u_\varepsilon, A_\varepsilon)$ *minimizes* $G_\varepsilon$*, then*

$$\frac{\tilde{\mu}(u_\varepsilon, A_\varepsilon)}{2\pi n_\varepsilon} \rightharpoonup \mu_0$$

*where* $\tilde{\mu}(u_\varepsilon, A_\varepsilon)$ *is the push-forward of the measure* $\mu(u_\varepsilon, A_\varepsilon)$ *under the blow-up* $x \mapsto \sqrt{\frac{h_{\text{ex}}}{n_\varepsilon}}(x - p)$*, and* $\mu_0$ *is the unique minimizer over probability measures of*

$$I(\mu) = -\pi \int_{\mathbb{R}^2\times\mathbb{R}^2} \log|x-y|\, d\mu(x)\, d\mu(y) + \pi \int_{\mathbb{R}^2} Q(x)\, d\mu(x). \tag{14}$$

Here, $n_\varepsilon$ corresponds to the expected optimal number of vortices. In [47] our result is really phrased as a $\Gamma$-convergence of $G_\varepsilon$ in the regime $1 \ll n \ll h_{\text{ex}}$, reducing the minimization of $G_\varepsilon$ to that of the limiting energy $I$. The problem of minimizing $I$ is a classical one in potential theory. Its minimizer $\mu_0$ is a probability measure of constant density over a subdomain of $\mathbb{R}^2$ (typically a disc or an ellipse). This result is in continuous connection with Theorem 3.4, except $n_\varepsilon \gg 1$. Again, vortices in the minimizers converge to $p$ as $\varepsilon \to 0$, and when one blows up at the right scale $\sqrt{\frac{h_{\text{ex}}}{n_\varepsilon}}$ around $p$, one obtains a uniform density of vortices $\mu_0$ in a subdomain of $\mathbb{R}^2$.

**3.2.3. Global minimizers in the regime $n_\varepsilon$ proportional to $h_{\text{ex}}$.** This happens in the next regime: $h_{\text{ex}} \sim \lambda|\log\varepsilon|$ with $\lambda > C(\Omega)$.

**Theorem 3.6** ([42], [47]). *Assume* $h_{\text{ex}} = \lambda|\log\varepsilon|$ *where* $\lambda > 0$ *is a constant independent of* $\varepsilon$*. As* $\varepsilon \to 0$*,* $\frac{G_\varepsilon}{h_{\text{ex}}^2}$ $\Gamma$*-converges to*

$$E_\lambda(\mu) = \frac{\|\mu\|}{2\lambda} + \frac{1}{2}\int_\Omega |\nabla h_\mu|^2 + |h_\mu - 1|^2,$$

*defined over bounded Radon measures which are in* $H^{-1}(\Omega)$*, where* $\|\mu\|$ *is the total mass of* $\mu$ *and* $h_\mu$ *is the solution to*

$$\begin{cases} -\Delta h_\mu + h_\mu = \mu & \text{in } \Omega, \\ h_\mu = 1 & \text{on } \partial\Omega. \end{cases} \tag{15}$$

*Consequently, if* $(u_\varepsilon, A_\varepsilon)$ *minimizes* $G_\varepsilon$*, then*

$$\frac{\mu(u_\varepsilon, A_\varepsilon)}{h_{\text{ex}}} \rightharpoonup \mu_*,$$

$\mu_*$ *being the unique minimizer over* $H^{-1}(\Omega) \cap (C_0^0(\Omega))^*$ *of* $E_\lambda$*.*

Observe also that

$$E_\lambda(\mu) = \frac{1}{2\lambda}\int_\Omega |\mu| + \frac{1}{2}\int_{\Omega\times\Omega} G(x,y)\, d(\mu-1)(x)\, d(\mu-1)(y) \qquad (16)$$

where $G$ is the solution to $-\Delta G + G = \delta_y$ with $G = 0$ on $\partial\Omega$. That way, the similarity with $I$ is more apparent.

Again, by $\Gamma$-convergence, we reduce to minimizing the limiting energy $E_\lambda$ on the space of bounded Radon measures on $\Omega$. It turns out that this problem is dual, in the sense of convex duality, to an obstacle problem:

**Proposition 3.7.** *$\mu$ minimizes $E_\lambda$ if and only if $h_\mu$ is the minimizer for*

$$\min_{\substack{h\geq 1-\frac{1}{2\lambda}\\ h=1 \text{ on } \partial\Omega}} \int_\Omega |\nabla h|^2 + h^2. \qquad (17)$$

The solution of the obstacle problem (17) is well-known, and given by a variational inequality (see [28]). Obstacle problems are a particular type of *free-boundary problems*, the free-boundary here being the boundary of the coincidence set

$$\omega_\lambda = \left\{ x\in\Omega / h(x) = 1 - \frac{1}{2\lambda} \right\}.$$

Then $h$ verifies $-\Delta h + h = 0$ outside of $\omega_\lambda$, so $\omega_\lambda$ is really the support of $\mu_*$, on which $\mu_*$ is equal to the constant density $\left(1-\frac{1}{2\lambda}\right)dx$. An easy analysis of this obstacle problem yields the following:

1. $\omega_\lambda = \varnothing$ (hence $\mu_* = 0$) if and only if $\lambda < C(\Omega)$, where $C(\Omega)$ was given by (11). (This corresponds to the case $h_{\text{ex}} < H_{c_1}$.)

2. For $\lambda = C(\Omega)$, $\omega_\lambda = \{p\}$. This is the case when $h_{\text{ex}} \sim H_{c_1}$ at leading order. In the scaling chosen here $\mu_* = 0$ but the true behavior of the vorticity is ambiguous unless going to the next order term as done in Theorems 3.4 and 3.5.

3. For $\lambda > C(\Omega)$, the measure of $\omega_\lambda$ is nonzero, so the limiting vortex density $\mu_* \neq 0$. Moreover, as $\lambda$ increases (i.e. as $h_{\text{ex}}$ does), $\omega_\lambda$ increases. When $\lambda = +\infty$, $\omega_\lambda$ becomes $\Omega$ and $\mu_* = 1$, this corresponds to the case $h_{\text{ex}} \gg |\log\varepsilon|$ of the next subsection.

**3.2.4. Global minimizers in the regime $|\log\varepsilon| \ll h_{\text{ex}} \ll \varepsilon^{-2}$.** For applied fields much larger than $|\log\varepsilon|$ but below $H_{c_2}$, even though the number of vortices becomes very large, the minimization problem becomes local and can be solved by blowing-up and using Theorem 3.6. The energy-density and the vortex repartition are thus found to be uniform, as seen in:

**Theorem 3.8** ([41], [47]). *Assume, as $\varepsilon \to 0$, that $|\log \varepsilon| \ll h_{\text{ex}} \ll 1/\varepsilon^2$. Then, if $(u_\varepsilon, A_\varepsilon)$ minimizes $G_\varepsilon$, and letting $g_\varepsilon(u, A)$ denote the energy-density $\frac{1}{2}\big(|\nabla_A u|^2 + |h - h_{\text{ex}}|^2 + \frac{1}{2\varepsilon^2}(1 - |u|^2)^2\big)$, we have*

$$\frac{2g_\varepsilon(u_\varepsilon, A_\varepsilon)}{h_{\text{ex}} \log \frac{1}{\varepsilon\sqrt{h_{\text{ex}}}}} \rightharpoonup dx \quad \text{as } \varepsilon \to 0$$

*in the weak sense of measures, where $dx$ denotes the two-dimensional Lebesgue measure; and thus*

$$\min_{(u,A)\in H^1\times H^1} G_\varepsilon(u, A) \sim \frac{|\Omega|}{2} h_{\text{ex}} \log \frac{1}{\varepsilon\sqrt{h_{\text{ex}}}} \quad \text{as } \varepsilon \to 0,$$

*where $|\Omega|$ is the area of $\Omega$. Moreover*

$$\frac{h_\varepsilon}{h_{\text{ex}}} \to 1 \quad \text{in } H^1(\Omega)$$
$$\frac{\mu(u_\varepsilon, A_\varepsilon)}{h_{\text{ex}}} \to dx \quad \text{in } H^{-1}(\Omega).$$

In Theorems 3.5, 3.6 and 3.8 we find an optimal limiting density which is constant on its support ($\omega_\lambda$ or $\Omega$). This provides a first (but very incomplete) confirmation of the Abrikosov lattices of vortices observed and predicted in physics.

**3.2.5. Global minimizers in higher applied field.** Here, we will present the situation with decreasing applied field. For large enough applied field, the only solution is the (trivial) normal one $u \equiv 0$, $h \equiv h_{\text{ex}}$.

Giorgi and Phillips have proved in [16] that this is the case for $h_{\text{ex}} \geq C\varepsilon^{-2}$.

**Theorem 3.9** (Giorgi–Phillips [16]). *There exists a constant $C$ such that if $h_{\text{ex}} \geq C\varepsilon^{-2}$ and $\varepsilon$ is small enough, then the only solution to* (GL) *is the normal one $u \equiv 0$, $h \equiv h_{\text{ex}}$.*

This result implies the upper bound $H_{c_3} \leq C\varepsilon^{-2}$ for that constant $C$.

Decreasing the applied field to $H_{c_3}$, a bifurcation from the normal solution of a branch of solutions with surface superconductivity occurs. The linear analysis of this bifurcation was first performed in the half-plane by De Gennes [13], then by Bauman–Phillips–Tang [4] in the case of a disc; and for general domains, formally by Chapman [11], Bernoff–Sternberg [5], Lu and Pan [33], then rigorously by Del Pino–Felmer–Sternberg [14], Helffer–Morame [19], Helffer–Pan [18]. The nucleation of surface superconductivity takes place near the point of maximal curvature of the boundary, and the asymptotics for $H_{c_3}$ is given by the following result.

**Theorem 3.10.**

$$H_{c_3} \sim \frac{\varepsilon^{-2}}{\beta_0} + \frac{C_1}{\beta_0^{3/2}} \max(\text{curv}(\partial\Omega))\varepsilon^{-1},$$

*where $\beta_0$ is the smallest eigenvalue of a Schrödinger operator with magnetic field in the half-plane.*

The behavior of energy minimizers for $H_{c_2} \leq h_{\text{ex}} \leq H_{c_3}$ has been studied by Pan [37] (see also Almog), who showed that, as known by physicists, minimizers present surface superconductivity which spreads to the whole boundary, with exponential decay of $|u|$ from the boundary of the domain.

At $H_{c_2}$, one goes from surface superconductivity to bulk-superconductivity. It was established by Pan [37] that

$$H_{c_2} = \varepsilon^{-2}.$$

Qualitative results on bulk-superconductivity below $H_{c_2}$ were obtained in [44], where we established in particular how bulk-superconductivity increases (in average) as $h_{\text{ex}}$ is lowered immediately below $H_{c_2}$.

## 4. Local minimizers: branches of solutions

The techniques developed to describe energy-minimizers also allow to find branches of locally minimizing (hence stable, and physically observable) solutions, with prescribed numbers of vortices. This is a problem of inverse type: given critical points or minimizers of the limiting energy $w$, can we find critical points / local minimizers of $G_\varepsilon$ which converge to it? For answers on that question regarding $E_\varepsilon$, see the book of Pacard and Rivière [36].

**Theorem 4.1** ([49], [47]). *For $\varepsilon < \varepsilon_0$, and for any $n$ and $h_{\text{ex}}$ belonging to appropriate intervals, there exists a locally minimizing critical point $(u_\varepsilon, A_\varepsilon)$ of $G_\varepsilon$ such that $u_\varepsilon$ has exactly $n$ zeroes $a_1^\varepsilon, \ldots, a_n^\varepsilon$ and there exists $R > 0$ such that $|u_\varepsilon| \geq \frac{1}{2}$ in $\Omega \backslash \bigcup_i B(a_i^\varepsilon, R\varepsilon)$, with $\deg(u_\varepsilon, \partial B(a_i^\varepsilon, R)) = 1$. Moreover:*

1. *If $n$ and $h_{\text{ex}}$ are constant, independent of $\varepsilon$, up to extraction of a subsequence, the configuration $(a_1^\varepsilon, \ldots, a_n^\varepsilon)$ converges as $\varepsilon \to 0$ to a minimizer of the function*

$$R_{n,h_{\text{ex}}} = -\pi \sum_{i \neq j} \log |x_i - x_j| + \pi \sum_{i,j} S_\Omega(x_i, x_j) + 2\pi h_{\text{ex}} \sum_{i=1}^{n} (h_0 - 1)(x_i).$$

   *where $S_\Omega$ is the regular part of a Green function associated to $\Omega$.*

2. *If $n = O(1)$ and $h_{\text{ex}} \to \infty$, up to extraction of a subsequence, the configuration of the $\tilde{a}_i^\varepsilon = \sqrt{\frac{h_{\text{ex}}}{n}}(a_i^\varepsilon - p)$ converges as $\varepsilon \to 0$ to a minimizer of $w_n$.*

3. *If $n \to \infty$ and $h_{\text{ex}} \to \infty$, then denoting again $\tilde{a}_i^\varepsilon = \sqrt{\frac{h_{\text{ex}}}{n}}(a_i^\varepsilon - p)$,*

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{\tilde{a}_i^\varepsilon} \rightharpoonup \mu_0$$

   *where $\mu_0$ is the unique minimizer of $I$ (defined in* (14)).

The method of the proof consists in finding these solutions as local minimizers by minimizing $G_\varepsilon$ over some open sets of the type $U_n = \{(u, A)/\pi(n-1)|\log\varepsilon| < G^0_\varepsilon(u, A) < \pi(n+1)|\log\varepsilon|\}$. Minimizing over $U_n$ consists, roughly speaking, in minimizing over configurations with $n$ vortices, the difficulty is in proving that the minimum over $U_n$ is achieved at an interior point (this comes from the quantization of the energetic cost of vortices), thus yielding a local energy minimizer.

We thus show the multiplicity of locally minimizing solutions, for a given $h_{\text{ex}}$, in a wide range (from $h_{\text{ex}} = O(1)$ to $h_{\text{ex}} \gg |\log\varepsilon|$): essentially, solutions with 0, 1, 2, 3, ... vortices coexist and are all stable, even if not energy-minimizing.

We also have derived multiple "renormalized energies" $R_{n,h_{\text{ex}}}$, $w_n$, $I(\mu)$ corresponding to the three regimes above. Observe that $w_n$ corresponds somewhat to the limit of $R_{n,h_{\text{ex}}}$ as $h_{\text{ex}} \to \infty$, while $I$ is a continuum limit as $n \to \infty$ (but still $n \ll h_{\text{ex}}$) of $w_n$. $E_\lambda$ can also be seen as the limit as both $n$ and $h_{\text{ex}}$ tend to $\infty$ but $n/h_{\text{ex}}$ not tending to 0. Thus these limiting or renormalized energies are not only valid for global minimization, but also for local minimization.

## 5. Critical points approach

The issue here is to derive conditions on limiting vortices or vortex-densities just assuming that we start from a family of solutions to (GL) or critical points of $G_\varepsilon$, not necessarily stable. This strategy was already implemented for the functional $E_\varepsilon$ in [6], leading to

**Theorem 5.1** (Bethuel–Brezis–Hélein [6]). *If $u_\varepsilon$ is a sequence of solutions of* (4) *in $\Omega$ with $u_\varepsilon = g$ on $\partial\Omega$, $\deg(g) = d > 0$, and $E_\varepsilon(u_\varepsilon) \le C|\log\varepsilon|$ then, as $\varepsilon \to 0$ and up to extraction of a subsequence, there exist distinct points $a_1, \dots, a_n \in \Omega$, and degrees $d_1, \dots, d_n$ with $\sum_{i=1}^n d_i = d$, such that $u_\varepsilon \to u_*$ in $C^k_{\text{loc}}\big(\Omega\backslash\bigcup_i\{a_i\}\big)$ where $u_*$ is a harmonic map from $\Omega\backslash\{a_1, \dots, a_n\}$ to $\mathbb{S}^1$ with $u_* = g$ on $\partial\Omega$ and with degree $d_i$ around each $a_i$. Moreover $(a_1, \dots, a_n)$ is a critical point of $W$ (the $d_i$'s being fixed).*

Thus, the vortices of critical points of $E_\varepsilon$ converge to critical points of the limiting energy $W$.

It is a corresponding result that we obtain for the vortex-densities for $G_\varepsilon$. The strategy consists similarly in passing to the limit $\varepsilon \to 0$, not in (GL), but in the stationarity relation

$$\frac{d}{dt}_{|t=0} G_\varepsilon(u \circ \chi_t, A \circ \chi_t) = 0$$

satisfied for the critical points (with $\chi_t$ a one-parameter family of diffeomorphisms such that $\chi_0 = Id$). That relation is equivalent by Noether's theorem to a relation of the form

$$\text{div } T_\varepsilon = 0$$

where $T_\varepsilon$ is called the "stress-energy" or "energy-momentum" tensor. For the present energy-functional

$$T_\varepsilon = \frac{1}{2}\begin{pmatrix} |\partial_1^A u|^2 - |\partial_2^A u|^2 & 2\langle \partial_1^A u, \partial_2^A u\rangle \\ 2\langle \partial_1^A u, \partial_2^A u\rangle & |\partial_2^A u|^2 - |\partial_1^A u|^2 \end{pmatrix} + \left(\frac{h^2}{2} - \frac{(1-|u|^2)^2}{2\varepsilon^2}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where $\partial_j^A = \partial_j - iA_j$.

In what follows we assume that $(u_\varepsilon, A_\varepsilon)$ are sequences of critical points of $G_\varepsilon$ such that $G_\varepsilon(u_\varepsilon, A_\varepsilon) \le C\varepsilon^{-\alpha}$, $\alpha < \frac{1}{3}$, and $n_\varepsilon$ is defined as $\sum_i |d_i|$ where the $d_i$'s are the degrees of the balls of total radius $r = \varepsilon^{2/3}$ given by Theorem 2.1.

**Theorem 5.2** ([43], [47]). *Let $(u_\varepsilon, A_\varepsilon)$ and $n_\varepsilon$ be as above. If $n_\varepsilon$ vanishes in a neighborhood of 0 then $\mu(u_\varepsilon, A_\varepsilon)$ tends to 0 in $W^{-1,p}(\Omega)$ for some $p \in (1,2)$. If not, then going to a subsequence*

$$\frac{\mu(u_\varepsilon, A_\varepsilon)}{n_\varepsilon} \to \mu \tag{18}$$

*in $W^{-1,p}(\Omega)$ for some $p \in (1,2)$ where $\mu$ is a measure. Moreover, one of the two following possibilities occur (after extraction of a subsequence if necessary).*

1. *If $n_\varepsilon = o(h_{\mathrm{ex}})$ then*
$$\mu \nabla h_0 = 0. \tag{19}$$
2. *If $h_{\mathrm{ex}}/n_\varepsilon \to \lambda \ge 0$, then, letting $h_\mu$ be the solution of (15), the symmetric 2-tensor $T_\mu$ with coefficients*
$$T_{ij} = \frac{1}{2}\begin{pmatrix} |\partial_1 h_\mu|^2 - |\partial_2 h_\mu|^2 & 2\partial_1 h_\mu \partial_2 h_\mu \\ 2\partial_1 h_\mu \partial_2 h_\mu & |\partial_2 h_\mu|^2 - |\partial_1 h_\mu|^2 \end{pmatrix} - \frac{h_\mu^2}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
*is divergence-free in finite part.*

*In the latter case, if $\mu$ is such that $h_\mu \in H^1(\Omega)$ then $T_\mu$ is in $L^1$ and divergence-free in the sense of distributions. Moreover $|\nabla h_\mu|^2$ is in $W^{1,q}_{\mathrm{loc}}(\Omega)$ for any $q \in [1, +\infty)$.*

*If we assume in addition that $\mu \in L^p(\Omega)$ for some $p > 1$, then*

$$\mu \nabla h_\mu = 0.$$

*Finally, if we assume $\nabla h_\mu \in C^0(\Omega) \cap W^{1,1}(\Omega)$ (this is the case if $\mu$ is in $L^p$, for some $p > 1$ for instance), then $h_\mu$ is in $C^{1,\alpha}(\Omega)$ for any $\alpha \in (0,1)$ and $0 \le h_\mu \le 1$. In this case*

$$\mu = h_\mu \mathbf{1}_{\{|\nabla h_\mu| = 0\}}, \tag{20}$$

*(where $\mathbf{1}$ stands for the characteristic function) and thus $\mu$ is a nonnegative $L^\infty$ function.*

To sum up, the limiting condition is $\mu\nabla h_0$ in the first case, it means that when there are too few of them, vortices all concentrate at the critical points of $h_0$ at the limit. In the second case, it is a weak form of the relation $\mu\nabla h_\mu = 0$ (which cannot be written as such when $h_\mu$ is not regular enough, counterexamples can be built).

We obtained an analogous result for critical points of $E_\varepsilon$ with possibly large numbers of vortices.

Also, once more, the result has a higher-dimensional version for the functional $E_\varepsilon$: as mentioned in Section 3.1.2, it was proved in [30], [7] that the vorticities for critical points of $E_\varepsilon$ converge to stationary varifolds, i.e. critical points for the length/area.

## 6. Study of the dynamics

The philosophy that has been successful in the minimization approach has been to extract limiting reduced energies (most often depending on some parameter regimes). These energies come up as $\Gamma$-limits, thus giving the behavior of energy-minimizers, but we have seen that they are not only relevant for energy-minimizers, but also for critical points ("critical points converge to critical points") and for local minimizers. It turns out that these limiting energies are also relevant for the study of dynamical problems, such as that of the heat-flow of (GL) and (4).

### 6.1. Energy-based method

**6.1.1. Abstract argument.** In [46], [50], we gave criteria to determine when a family of energies $F_\varepsilon$ converges to its $\Gamma$-limit $F$ in a sort of $C^1$ or $C^2$ sense which allows to pass to the limit in the associated gradient-flow (we called this "$\Gamma$-convergence of gradient flows"). The abstract situation is the following: assume that $F_\varepsilon$ $\Gamma$-converges to $F$ for the sense of convergence $S$ (the sense of convergence can be a weak convergence or a convergence of some nonlinear quantity), that means in particular

$$\underline{\lim}_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \geq F(u) \quad \text{when } u_\varepsilon \overset{S}{\rightharpoonup} u, \tag{21}$$

and consider the gradient of $F_\varepsilon$ with respect to some Hilbert structure $X_\varepsilon$, denoted $\nabla_{X_\varepsilon} F_\varepsilon$. The question is to find conditions to get that solutions of the gradient flow $\partial_t u_\varepsilon = -\nabla_{X_\varepsilon} F_\varepsilon(u_\varepsilon)$ converge (in the sense $S$) to a solution of the gradient flow of $F$ with respect to some structure $Y$ to be determined. In the problem on Ginzburg–Landau vortices, $F_\varepsilon$ should be taken to be $E_\varepsilon - \pi n|\log\varepsilon|$ and $F = W$ (see Proposition 3.2), the sense of convergence to be considered is $u_\varepsilon \rightharpoonup^S u = (a_1, \ldots, a_n)$ if $\operatorname{curl}\langle iu_\varepsilon, \nabla u_\varepsilon\rangle \rightharpoonup 2\pi\sum_{i=1}^n d_i\delta_{a_i}$, where the $d_i = \pm 1$ and $n$ are fixed a priori. In that case the limiting flow is a finite dimensional one, so the proof of existence of its solution is easy.

The two sufficient abstract conditions are that there should exist another Hilbert structure $Y$ on the space where $F$ is defined, satisfying the following:

1) For a subsequence such that $u_\varepsilon(t) \overset{S}{\rightharpoonup} u(t)$ for every $t \in [0, T)$, we have for all $s \in [0, T)$,
$$\varliminf_{\varepsilon\to 0} \int_0^s \|\partial_t u_\varepsilon\|^2_{X_\varepsilon}(t)\,dt \geq \int_0^s \|\partial_t u\|^2_Y(t)\,dt. \tag{22}$$

2) For any $u_\varepsilon \overset{S}{\rightharpoonup} u$
$$\varliminf_{\varepsilon\to 0} \|\nabla_{X_\varepsilon} F_\varepsilon(u_\varepsilon\|^2_{X_\varepsilon} \geq \|\nabla_Y F(u)\|^2_Y. \tag{23}$$

These conditions suffice in the case where $F$ is defined on a finite-dimensional space, to derive that if $u_\varepsilon$ solves
$$\partial_t u_\varepsilon = -\nabla_{X_\varepsilon} F_\varepsilon(u_\varepsilon) \quad \text{on } [0, T) \tag{24}$$
with $u_\varepsilon(0) \overset{S}{\rightharpoonup} u_0$, and is well-prepared in the sense that $F_\varepsilon(u_\varepsilon(0)) = F(u_0) + o(1)$, then $u_\varepsilon(t) \overset{S}{\rightharpoonup} u(t)$, where $u$ is the solution to
$$\begin{cases} \partial_t u = -\nabla_Y F(u), \\ u(0) = u_0. \end{cases}$$

The proof of this abstract result is rather elementary: for all $t < T$ we may write
$$\begin{aligned} F_\varepsilon(u_\varepsilon(0)) - F_\varepsilon(u_\varepsilon(t)) &= -\int_0^t \langle \nabla_{X_\varepsilon} F_\varepsilon(u_\varepsilon(s)),\, \partial_t\, u_\varepsilon(s)\rangle_{X_\varepsilon}\, ds \\ &= \frac{1}{2}\int_0^t \|\nabla_{X_\varepsilon} F_\varepsilon(u_\varepsilon)\|^2_{X_\varepsilon} + \|\partial_t u_\varepsilon\|^2_{X_\varepsilon}\, ds \\ &\geq \frac{1}{2}\int_0^t \|\nabla_Y F(u)\|^2_Y + \|\partial_t u\|^2_Y\, ds - o(1) \\ &\geq \int_0^t -\langle \nabla_Y F(u(s)),\, \partial_t u(s)\rangle_Y\, ds - o(1) \qquad (25) \\ &= F(u(0)) - F(u(t)) - o(1), \end{aligned}$$

hence
$$F(u(0)) - F(u(t)) \leq F_\varepsilon(u_\varepsilon(0)) - F_\varepsilon(u_\varepsilon(t)) + o(1).$$
But by well-preparedness, $F_\varepsilon(u_\varepsilon(0)) = F(u(0)) + o(1)$, thus
$$F_\varepsilon(u_\varepsilon(t)) \leq F(u(t)) + o(1).$$
But $F_\varepsilon \xrightarrow{\Gamma} F$ implies $\varliminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon(t)) \geq F(u(t))$. Therefore we must have equality everywhere and in particular equality in the Cauchy–Schwarz type relation (25), that is
$$\frac{1}{2}\int_0^t \|\nabla_Y F(u)\|^2_Y + \|\partial_t u\|^2_Y\, ds = \int_0^t -\langle \nabla_Y F(u(s)),\, \partial_t u(s)\rangle_Y$$

or

$$\int_0^t \|\nabla F(u) + \partial_t u\|_Y^2 \, ds = 0.$$

Hence, we conclude that $\partial_t u = -\nabla_Y F(u)$, for a.e. $t \in [0, T)$.

The method should and can be extended to infinite-dimensional limiting spaces and to the case where the Hilbert structures $X_\varepsilon$ and $Y$ (in particular $Y$) depend on the point, forming a sort of Hilbert manifold structure. In fact we can write down an analogous abstract result using the theory of "minimizing movements" of De Giorgi formalized by Ambrosio–Gigli–Savarè [3], a notion of gradient flows on structures which are not differentiable but simply metric structures.

**6.1.2. The result for Ginzburg–Landau without magnetic field.** Applying this abstract method to $F_\varepsilon = E_\varepsilon - \pi n |\log \varepsilon|$ and $F = W$ (with a prescribed number of vortices, and prescribed degrees), we retrieve the dynamical law of vortices which had been first established by Lin and Jerrard–Soner by PDE methods:

**Theorem 6.1** ([29], [25], [46]). *Let $u_\varepsilon$ be a family of solutions of*

$$\frac{\partial_t u}{|\log \varepsilon|} = \Delta u + \frac{u}{\varepsilon^2}(1 - |u|^2) \quad \text{in } \Omega$$

*with either*

$$u_\varepsilon = g \ \text{ on } \partial\Omega \quad \text{or} \quad \frac{\partial u_\varepsilon}{\partial n} = 0 \ \text{ on } \partial\Omega$$

*such that*

$$\operatorname{curl} \langle i u_\varepsilon, \nabla u_\varepsilon \rangle (0) \rightharpoonup 2\pi \sum_{i=1}^n d_i \delta_{a_i^0} \quad \text{as } \varepsilon \to 0$$

*with $a_i^0$ distinct points in $\Omega$, $d_i = \pm 1$, and*

$$E_\varepsilon(u_\varepsilon)(0) - \pi n |\log \varepsilon| \le W(a_i^0) + o(1). \tag{26}$$

*Then there exists $T^* > 0$ such that for all $t \in [0, T^*)$,*

$$\operatorname{curl} \langle i u, \nabla u \rangle (t) \rightharpoonup 2\pi \sum_{i=1}^n d_i \delta_{a_i(t)}$$

*as $\varepsilon \to 0$, with*

$$\begin{cases} \dfrac{da_i}{dt} = -\dfrac{1}{\pi} \nabla_i W(a_1(t), \dots, a_n(t)), \\ a_i(0) = a_i^0, \end{cases} \tag{27}$$

*where $T^*$ is the minimum of the collision time and exit time of the vortices under this law.*

Thus, as expected, vortices move along the gradient flow for their interaction $W$, and this reduces the PDE to a finite dimensional evolution (a system of ODE's).

The difficulty is in proving that the abstract conditions (22)–(23) hold in the Ginzburg–Landau setting. For example the first relation (22) relates the velocity of underlying vortices to $\partial_t u_\varepsilon$ and can be read

$$\varliminf_{\varepsilon\to 0} \frac{1}{|\log\varepsilon|} \int_{[0,t]\times\Omega} |\partial_t u_\varepsilon|^2 \, ds \geq \pi \sum_i \int_0^t |d_t\, a_i|^2 \, ds \tag{28}$$

assuming $\operatorname{curl}\langle iu, \nabla u_\varepsilon\rangle(t) \rightharpoonup 2\pi \sum_i d_i \delta_{a_i(t)}$, as $\varepsilon \to 0$, for all $t$ and $d_i = \pm 1$. This turns out to hold as a general relation (as proved in [45]), without requiring the configurations to solve any particular equation; it is related to the topological nature of the vortices.

**6.1.3. Dynamical law for Ginzburg–Landau with magnetic field.** By the same method, we obtained the dynamics of a bounded number of vortices for the full Ginzburg–Landau equations with magnetic field, i.e. the gradient-flow of (1), for large applied fields (the result for bounded applied fields had been obtained by Spirn [52]).

Assuming that $h_{\text{ex}} = \lambda|\log\varepsilon|$, $0 < \lambda < \infty$, we obtained in [46] that, for energetically well-prepared solutions $(u_\varepsilon, A_\varepsilon)$ of (GL), such that $\mu(u_\varepsilon, A_\varepsilon)(0) \rightharpoonup 2\pi \sum_{i=1}^n d_i \delta_{a_i^0}$, with $d_i = \pm 1$, we have for all $t \in [0, T^*)$,

$$\mu(u_\varepsilon, A_\varepsilon)(t) \rightharpoonup 2\pi \sum_{i=1}^n d_i \delta_{a_i(t)}$$

with the dynamical law

$$\frac{da_i}{dt} = -d_i \lambda \nabla h_0(a_i(t)), \quad a_i(0) = a_i^0$$

for all $i$, where $T^*$ is the minimum of the collision time and of the exit time from $\Omega$ for this law of motion.

**6.1.4. Stability of critical points.** In [50], we extended the "$\Gamma$-convergence of gradient flows" method above to the second order, i.e. we gave conditions on the second derivatives of the energies $F_\varepsilon$ $\Gamma$-converging to $F$ which ensure that critical points of $F_\varepsilon$ converge to critical points of $F$ (condition (23) above already ensures it) and that *moreover* stable critical points (in the sense of nonnegative Hessian) of $F_\varepsilon$ converge to stable critical points of $F$ (and more generally bounding from below the Morse index of the critical points of $F_\varepsilon$ by that of those of $F$). The abstract condition is roughly the following: for any family $u_\varepsilon$ of critical points of $F_\varepsilon$ such that $u_\varepsilon \overset{S}{\rightharpoonup} u$; for any $V$, we can find $v_\varepsilon(t)$ defined in a neighborhood of $t = 0$, such that $\partial_t v_\varepsilon(0)$

depends on $V$ in a linear and one-to-one manner, and

$$v_\varepsilon(0) = u_\varepsilon \tag{29}$$

$$\lim_{\varepsilon\to 0} \tfrac{d}{dt}_{|t=0} F_\varepsilon(v_\varepsilon(t)) = \tfrac{d}{dt}_{|t=0} F(u+tV) = dF(u).V \tag{30}$$

$$\lim_{\varepsilon\to 0} \tfrac{d^2}{dt^2}_{|t=0} F_\varepsilon(v_\varepsilon(t)) = \tfrac{d^2}{dt^2}_{|t=0} F(u+tV) = \langle D^2F(u)V, V\rangle. \tag{31}$$

We show that these conditions hold for (3) and deduce the result

**Theorem 6.2** ([50]). *Let $u_\varepsilon$ be a family of solutions of* (4) *such that $E_\varepsilon(u_\varepsilon) \le C|\log\varepsilon|$, with either Dirichlet or homogeneous Neumann boundary conditions. Then, there exists a family of points $a_1, \dots, a_n$ and nonzero integers $d_1, \dots, d_n$ such that, up to extraction of a subsequence,*

$$\operatorname{curl}(iu_\varepsilon, \nabla u_\varepsilon) \rightharpoonup 2\pi \sum_{i=1}^{n} d_i \delta_{a_i},$$

*where $(a_1, \dots, a_n)$ is a critical point of $W$. Moreover, if $u_\varepsilon$ is a stable solution of* (4) *then $(a_1, \dots, a_n)$ is a stable critical point of $W$; and more generally, denoting by $n_\varepsilon^+$ the dimension of the space spanned by eigenvectors of $D^2E_\varepsilon(u_\varepsilon)$ associated to positive eigenvalues and $n^+$ the dimension of the space spanned by eigenvectors of $D^2W(a_i)$ associated to positive eigenvalues* (*resp. $n_\varepsilon^-$ and $n^-$ for negative eigenvalues*)*, we have, for $\varepsilon$ small enough,*

$$n_\varepsilon^+ \ge n^+, \quad n_\varepsilon^- \ge n^-. \tag{32}$$

One of the interesting consequences of the theorem is the following, a consequence of the fact that the renormalized energy in Neumann boundary condition has no nontrivial stable critical point.

**Theorem 6.3.** *Let $u_\varepsilon$ be a family of nonconstant solutions of* (4) *with homogeneous Neumann boundary condition $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega$, such that $E_\varepsilon(u_\varepsilon) \le C|\log\varepsilon|$, then for $\varepsilon$ small enough, $u_\varepsilon$ is unstable.*

This shows that the model without magnetic field (4) cannot stabilize vortices, contrarily to the one with nonzero applied magnetic field.

This extended a result of Jimbo–Morita [26] (see also Jimbo–Sternberg [27] with magnetic field) valid for any $\varepsilon$ but for convex domains.

**6.2. PDE-based results.** Most results on convergence of solutions of the Ginzburg–Landau flow to solutions of the flow for limiting energies were proved by PDE-based methods. We briefly review them.

**6.2.1. Heat flow in higher dimensions.** The convergence of the flow for $E_\varepsilon$ is also true in higher dimensions where the limiting energy-density is length/surface, it was established (see [8], [31]) that the limit of the parabolic evolution of $E_\varepsilon$ is a Brakke flow (a weak form of mean-curvature flow, which is the expected gradient flow of the limiting energy).

**6.2.2. Other flows.** The Schrödinger flow of (4), also called the Gross–Pitaevskii equation, is considered in superfluids, nonlinear optics and Bose–Einstein condensation. The limiting dynamical law of vortices is still the corresponding one (i.e. the Hamiltonian flow) for the limiting renormalized energy

$$\frac{da_i}{dt} = -\frac{1}{\pi}\nabla_i^{\perp} W(a_1, \ldots, a_n).$$

The convergence was proved, still with well-prepared assumptions, by Colliander–Jerrard [12] on a torus, and by Lin–Xin [32] in the whole plane. In the case of the wave flow, the analogous limiting dynamical law was established by Jerrard in [23].

**6.3. Collision issues.** The result of Theorem 6.1 is valid only up to collision time under the law (27), but collisions do happen if there are vortices of opposite degrees. The question of understanding the collisions and extending the motion law passed them is delicate. Bethuel–Orlandi–Smets [9], [10] treated this question, as well as other issues of non well-prepared data, vortex-splitting and phase-vortex interaction in infinite domains.

In [51], the collision problem was approached with the idea of basing the study on the energy, like for the "Γ-convergence of gradient flow". We proved that when several vortices become very close to each other (but not too close) a dynamical law after blow-up can be derived through the same method presented above. When vortices become too close to apply this, we focused on evaluating energy dissipation rates, through the study of the perturbed Ginzburg–Landau equation

$$\Delta u + \frac{1}{\varepsilon^2} u(1 - |u|^2) = f_\varepsilon \quad \text{in } \Omega, \tag{33}$$

with Dirichlet or Neumann boundary data, where $f_\varepsilon$ is given in $L^2(\Omega)$ (the instantaneous energy-dissipation rate in the dynamics is exactly $|\log\varepsilon| \|f_\varepsilon\|^2_{L^2(\Omega)}$). We prove that the "energy-excess" (meaning the difference between $E_\varepsilon - \pi n |\log\varepsilon|$ and the renormalized energy $W$ of the underlying vortices) is essentially controlled by $C\|f_\varepsilon\|^2_{L^2}$. We then show that when $u$ solves (33) and has vortices which become very close, forming what we call an "unbalanced cluster" in the sense that $\sum_i d_i^2 \neq \left(\sum_i d_i\right)^2$ in the cluster (see [51] for a precise definition), then a lower bound for $\|f_\varepsilon\|_{L^2}$ must hold:

**Theorem 6.4** ([51]). *Let $u_\varepsilon$ solve* (33) *with $E_\varepsilon(u_\varepsilon) \leq C|\log\varepsilon|$, $|\nabla u_\varepsilon| \leq \frac{M}{\varepsilon}$ and $|u_\varepsilon| \leq 1$. There exists $l_0 > 0$ such that, if $u_\varepsilon$ has an unbalanced cluster of vortices at the scale $l < l_0$ then*

$$\|f_\varepsilon\|^2_{L^2(\Omega)} \geq \min\left(\frac{C}{l^2|\log\varepsilon|}, \frac{C}{l^2\log^2 l}\right). \tag{34}$$

In particular, when vortices get close to each other, say two vortices of opposite degrees for example, then they form an unbalanced cluster of vortices at scale

$l =$ their distance, and the relation (34) gives a large energy-dissipation rate (scaling like $1/l^2$). This serves to show that such a situation cannot persist for a long time and we are able to prove that the vortices collide and disappear in time $Cl^2 + o(1)$, with all energy-excess dissipating in that time. Thus after this time $o(1)$, the configuration is again "well-prepared" and Theorem 6.1 can be applied again, yielding the dynamical law with the remaining vortices, until the next collision, etc.

## References


[1] Abrikosov, A., On the magnetic properties of superconductors of the second type. *Soviet Phys. JETP* **5** (1957), 1174–1182.

[2] Alberti, G., Baldo, S., Orlandi, G., Functions with prescribed singularities. *J. Eur. Math. Soc.* (*JEMS*) **5** (3) (2003), 275–311.

[3] Ambrosio, L., Gigli, N., Savarè, G., *Gradient flows in metric spaces and in the Wasserstein spaces of probability measures*. Lectures Math. ETH Zurich, Birkhäuser, Basel 2005.

[4] Bauman, P., Phillips, D., Tang, Q., Stable nucleation for the Ginzburg-Landau system with an applied magnetic field. *Arch. Ration. Mech. Anal.* **142** (1) (1998), 1–43.

[5] Bernoff, A., Sternberg, P., Onset of superconductivity in decreasing fields for general domains, *J. Math. Phys.* **39** (3) (1998), 1272–1284.

[6] Bethuel, F., Brezis, H., Hélein, F., *Ginzburg-Landau vortices.* Progr. Nonlinear Differential Equations Appl. 13, Birkhäuser, Boston, MA, 1994.

[7] Bethuel, F., Brezis, H., Orlandi, G., Asymptotics for the Ginzburg-Landau equation in arbitrary dimensions. *J. Funct. Anal.* **186** (2) (2001), 432–520.

[8] Bethuel, F., Orlandi, G., Smets, D., Convergence of the parabolic Ginzburg-Landau equation to motion by mean curvature. *Ann. of Math.* **163** (1) (2006), 37–163.

[9] Bethuel, F., Orlandi, G., Smets, D., Collisions and phase-vortex interactions in dissipative Ginzburg-Landau dynamics. *Duke Math. J.*, to appear.

[10] Bethuel, F., Orlandi, G., Smets, D., Quantization and motion laws for Ginzburg-Landau vortices. Preprint, 2005.

[11] Chapman, S. J., Nucleation of superconductivity in decreasing fields. I, II. *Eur. J. Appl. Math.* **5** (1994), 449–468, 468–494.

[12] Colliander, J., Jerrard, R., Vortex dynamics for the Ginzburg-Landau-Schrödinger equation. *Internat. Math. Res. Notices* **1998** (7) (1998), 333–358.

[13] DeGennes, P. G., *Superconductivity of metal and alloys*. Benjamin, New York, Amsterdam 1966.

[14] Del Pino, M., Felmer, P., Sternberg, P., Boundary concentration for eigenvalue problems related to the onset of superconductivity. *Comm. Math. Phys.* **210** (2000), 413–446.

[15] Ginzburg, V. L., Landau, L. D., On the Theory of Superconductivity. In *Collected papers of L. D. Landau*. Ed. by D. ter Haar, Pergamon Press, New York 1965, 546–568.

[16] Giorgi, T., Phillips, D., The breakdown of superconductivity due to strong fields for the Ginzburg-Landau model. *SIAM J. Math. Anal.* **30** (2) (1999), 341–359.

[17] Gueron, S., Shafrir, I., On a Discrete Variational Problem Involving Interacting Particles. *SIAM J. Appl. Math.* **60** (1) (2000), 1–17.

[18] Helffer, B., Pan, X.-B., Upper critical field and location of surface nucleation of superconductivity. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **20** (1) (2003), 145–181.

[19] Helffer, B., Morame, A., Magnetic bottles in connection with superconductivity. *J. Funct. Anal.* **185** (2) (2001), 604–680.

[20] Helffer, B., Morame, A., Magnetic bottles for the Neumann problem: curvature effects in the case of dimension 3 (general case). *Ann. Sci. École Norm. Sup.* (4) **37** (1) (2004), 105–170.

[21] Jaffe, A., Taubes, C., *Vortices and monopoles*. Progr. Phys. 2, Birkhäuser, Boston 1980.

[22] Jerrard, R. L., Lower bounds for generalized Ginzburg-Landau functionals. *SIAM J. Math. Anal.* **30** (4) (1999), 721-746.

[23] Jerrard, R. L., Vortex dynamics for the Ginzburg-Landau wave equation. *Calc. Var. Partial Differential Equations* **9** (1) (1999), 1–30.

[24] Jerrard, R. L., Soner, H. M., The Jacobian and the Ginzburg-Landau energy. *Calc. Var. Partial Differential Equations* **14** (2) (2002), 151–191.

[25] Jerrard, R. L., Soner, H. M., Dynamics of Ginzburg-Landau vortices. *Arch. Ration. Mech. Anal.* **142** (2) (1998), 99–125.

[26] Jimbo, S., Morita, Y., Stable solutions with zeros to the Ginzburg-Landau equation with Neumann boundary condition. *J. Differential Equations* **128** (2) (1996), 596–613.

[27] Jimbo, S., Sternberg, P., Nonexistence of permanent currents in convex planar samples. *SIAM J. Math. Anal.* **33** (6) (2002), 1379–1392.

[28] Kinderlehrer, D., Stampacchia, G., An introduction to variational inequalities and their applications. Pure Appl. Math. 88, Academic Press, New York 1980.

[29] Lin, F.-H., Some dynamic properties of Ginzburg-Landau vortices. *Comm. Pure Appl. Math* **49** (1996), 323–359.

[30] Lin, F.-H., Rivière, T., A quantization property for static Ginzburg-Landau vortices. *Comm. Pure Appl. Math.* **54** (2) (2001), 206–228.

[31] Lin, F.-H., Rivière, T., A quantization property for moving line vortices. *Comm. Pure Appl. Math.* **54** (7) (2001), 826–850.

[32] Lin, F.-H., Xin, J.-X., On the dynamical law of the Ginzburg-Landau vortices on the plane. *Comm. Pure Appl. Math.* **52** (10) (1999), 1189–1212.

[33] Lu, K., Pan, X.-B., Estimates of the upper critical field for the Ginzburg-Landau equations of superconductivity. *Phys. D* **127** (1–2) (1999), 73–104.

[34] Manton, N., Sutcliffe, P., *Topological solitons*. Cambridge monographs on mathematical physics. Cambridge University Press, Cambridge 2004.

[35] Mironescu, P., Les minimiseurs locaux pour l'équation de Ginzburg-Landau sont à symétrie radiale. *C. R. Acad. Sci. Paris, Ser. I* **323** (6) (1996), 593–598.

[36] Pacard, F., Rivière, T., *Linear and nonlinear aspects of vortices*. Progr. Nonlinear Differential Equations Appl. 39, Birkhäuser Boston, Boston, MA, 2000.

[37] Pan, X.-B., Surface superconductivity in applied magnetic fields above $H_{c_2}$. *Comm. Math. Phys.* **228** (2) (2002), 327–370.

[38] Sandier, E., Lower bounds for the energy of unit vector fields and applications. *J. Funct. Anal.* **152** (2) (1998), 379–403; Erratum, Ibid. **171** (1) (2000), 233.

[39] Sandier, E., Ginzburg-Landau minimizers from $\mathbb{R}^{N+1}$ to $\mathbb{R}^N$ and minimal connections. *Indiana Univ. Math. J.* **50** (4) (2001), 1807–1844.

[40] Sandier, E., Serfaty, S., Global minimizers for the Ginzburg-Landau functional below the first critical magnetic field. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **17** (1) (2000), 119–145.

[41] Sandier, E., Serfaty, S., On the energy of type-II superconductors in the mixed phase. *Rev. Math. Phys.* **12** (9) (2000), 1219–1257.

[42] Sandier, E., Serfaty, S., A rigorous derivation of a free-boundary problem arising in superconductivity. *Ann. Sci. École Norm. Sup.* (4) **33** (4) (2000), 561–592.

[43] Sandier, E., Serfaty, S., Limiting vorticities for the Ginzburg-Landau equations. *Duke Math. Jour.* **117** (3) (2003), 403-446.

[44] Sandier, E., Serfaty, S., The decrease of bulk-superconductivity close to the second critical field in the Ginzburg-Landau model. *SIAM J. Math. Anal.* **34** (4) (2003), 939–956.

[45] Sandier, E., Serfaty, S., A product-estimate for Ginzburg-Landau and corollaries. *J. Funct. Anal.* **211** (1) (2004), 219–244.

[46] Sandier, E., Serfaty, S., Gamma-convergence of gradient flows with applications to Ginzburg-Landau. *Comm. Pure Appl. Math.* **57** (12) (2004), 1627–1672.

[47] Sandier, E., Serfaty, S., *Vortices in the Magnetic Ginzburg-Landau Model*. Birkhäuser, to appear.

[48] Serfaty, S., Local minimizers for the Ginzburg-Landau energy near critical magnetic field, I, II. *Commun. Contemp. Math.* **1** (1999), 213–254, 295–333.

[49] Serfaty, S., Stable configurations in superconductivity: uniqueness, multiplicity and vortex-nucleation. *Arch. Ration. Mech. Anal.* **149** (1999), 329–365.

[50] Serfaty, S., Stability in 2D Ginzburg-Landau passes to the limit. *Indiana Univ. Math. J.* **54** (1) (2005), 199–222.

[51] Serfaty, S., Vortex-collision and energy dissipation rates in the Ginzburg-Landau heat flow, part I: Study of the perturbed Ginzburg-Landau equation; part II: The dynamics. Submitted.

[52] Spirn, D., Vortex dynamics of the full time-dependent Ginzburg-Landau equations. *Comm. Pure Appl. Math.* **55** (5) (2002), 537–581.

[53] Tinkham, M., *Introduction to superconductivity*. Second edition. McGraw-Hill, New York 1996.

Courant Institute of Mathematical Sciences, 251 Mercer St, New York, NY 10012, U.S.A.
E-mail: serfaty@cims.nyu.edu

# Recent developments in elliptic partial differential equations of Monge–Ampère type

Neil S. Trudinger*

**Abstract.** In conjunction with applications to optimal transportation and conformal geometry, there has been considerable research activity in recent years devoted to fully nonlinear, elliptic second order partial differential equations of a particular form, given by functions of the Hessian plus a lower order matrix function. Regularity is determined through the behaviour of this function with respect to the gradient variables. We present a selection of second derivative estimates and indicate briefly their application to optimal transportation and conformal deformation of Riemannian manifolds.

**Mathematics Subject Classification (2000).** Primary 35J60, 35J65; Secondary 53A30.

**Keywords.** Fully nonlinear elliptic partial differential equations, Monge–Ampère type, optimal transportation, conformal deformation.

## 1. Introduction

In conjunction with applications to optimal transportation and conformal geometry, there has been considerable research activity in recent years devoted to fully nonlinear, elliptic second order partial differential equations of the form,

$$\mathcal{F}[u] := F\{D^2u + A(\cdot, u, Du)\} = B(\cdot, u, Du), \tag{1.1}$$

in domains $\Omega$ in Euclidean $n$-space, $\mathbb{R}^n$, as well as their extensions to Riemannian manifolds. Here the functions $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $A\colon \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$, $B\colon \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ are given and the resultant operator $\mathcal{F}$ is well-defined classically for functions $u \in C^2(\Omega)$. As customary $Du$ and $D^2u$ denote respectively the gradient vector and Hessian matrix of second derivatives of $u$, while we also use $x$, $z$, $p$, $r$ to denote points in $\Omega$, $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^n \times \mathbb{R}^n$ respectively with corresponding partial derivatives denoted, when there is no ambiguity, by subscripts. For example $F_r = \left[\frac{\partial F}{\partial r_{ij}}\right]$, $F_p = \left(\frac{\partial F}{\partial p_1}, \ldots, \frac{\partial F}{\partial p_n}\right)$ etc. The operator $\mathcal{F}$ is *elliptic* with respect to $u$ whenever the matrix

$$F_r\left\{D^2u + A(\cdot, u, Du)\right\} > 0. \tag{1.2}$$

Unless indicated otherwise we will assume the matrix $A$ is symmetric, but it is also important to address the possibility that it is not. When $A \equiv 0$, (1.1) reduces to the

*Research supported by the Australian Research Council.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

well-studied Hessian equation,

$$\mathcal{F}[u] = F\left(D^2u\right) = B, \tag{1.3}$$

while for $F(r) = \det r$, we obtain a *Monge–Ampère equation* of the form

$$\mathcal{F}[u] = \det\left\{D^2u + A(\cdot, u, Du)\right\} = B(\cdot, u, Du), \tag{1.4}$$

which is preserved under coordinate changes, unlike the standard Monge–Ampère equation, when $A \equiv 0$. The operator $\mathcal{F}$ in (1.4) is elliptic with respect to $u$ whenever

$$D^2u + A(\cdot, u, Du) > 0, \tag{1.5}$$

which implies $B > 0$.

Monge–Ampère equations of the general form (1.4) arise in applications, notably in optimal transportation, through the prescription of the absolute value of the Jacobian determinant of a mapping $T = T_u : \Omega \to \mathbb{R}^n$ given by

$$T_u = Y(\cdot, u, Du), \tag{1.6}$$

where $Y : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is a given vector field, that is

$$|\det DT_u| = \psi(\cdot, u, Du) \tag{1.7}$$

for a given nonnegative $\psi : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$. To write (1.7) in the form (1.4) we assume that $Y$ is differentiable and

$$\det Y_p \neq 0, \tag{1.8}$$

whence by calculation, we obtain

$$\mathcal{F}[u] = \det\{D^2u + Y_p^{-1}(Y_x + Y_z \otimes Du)\} = \psi/|\det Y_p|, \tag{1.9}$$

assuming $\mathcal{F}$ elliptic with respect to $u$. When the vector field $Y$ is independent of $z$ and generated by a *cost function* $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, through the equations

$$c_x(\cdot, Y(\cdot, p)) = p \tag{1.10}$$

we obtain the *optimal transportation* equation,

$$\mathcal{F}[u] = \det\left\{D^2u - D_x^2c(\cdot, Y(\cdot, Du))\right\} = \psi/|\det Y_p|. \tag{1.11}$$

Note that by differentiation of (1.10) we have

$$\begin{aligned} c_{x,y}(\cdot, Y) &= Y_p^{-1}, \\ c_{xx}(\cdot, Y) &= -Y_p^{-1}Y_x. \end{aligned} \tag{1.12}$$

In conformal geometry, equations of the form (1.1) arise from the study of the Schouten tensor of a Riemannian manifold under conformal deformation of its metric.

When the functions $F$ are homogeneous we obtain, in the special case of Euclidean space $\mathbb{R}^n$, equations of the form (1.1), where the matrix $A$ is given by

$$A(p) = -\tfrac{1}{2}|p|^2 I + p \otimes p. \tag{1.13}$$

Here we observe that, as in equation (1.9), we may write

$$A(p) = Y_p^{-1} Y_x(\cdot, p), \tag{1.14}$$

where $Y$ is the vector field

$$Y(x, p) = x - p/|p|^2 \tag{1.15}$$

generated by the cost function

$$c(x, y) = \log|x - y|. \tag{1.16}$$

The regularity of solutions of equations of the form (1.1) depends on the behaviour of the matrix $A$ with respect to the $p$ variables. Letting $\mathcal{U} \subset \Omega \times \mathbb{R} \times \mathbb{R}^n$, we say that $A$ is *regular* in $\mathcal{U}$ if

$$D_{p_k p_\ell} A_{ij}\, \xi_i\, \xi_j\, \eta_k\, \eta_\ell \leq 0 \tag{1.17}$$

in $\mathcal{U}$, for all $\xi, \eta \in \mathbb{R}^n$ with $\xi \cdot \eta = 0$; and *strictly regular* in $\mathcal{U}$ if there exists a constant $a_0 > 0$ such that

$$D_{p_k p_\ell} A_{ij}\, \xi_i\, \xi_j\, \eta_k\, \eta_\ell \leq -a_0|\xi|^2|\eta|^2 \tag{1.18}$$

in $\mathcal{U}$, for all $\xi, \eta \in \mathbb{R}^n$ with $\xi \cdot \eta = 0$.

These conditions were introduced in [23], [31] and called there A3w, A3 respectively. As we will explain in this paper, they are the natural conditions for regularity. Note that the matrix $A$ in (1.13) trivially satisfies (1.18) in $\mathbb{R}^n$, with $a_0 = 1$.

## 2. Second derivative estimates

The key estimates for classical solutions of equations of the form (1.1) are bounds for second derivatives as higher order estimates and regularity follows from the fully nonlinear theory [10], [21]. Here we present a selection of estimates for regular and strictly regular matrix functions $A$. For all these estimates we will assume $A$ and $B$ are $C^2$ smooth.

**2.1. Interior estimates.** Under the hypothesis of strict regularity we get quite strong interior estimates for very general functions $F$. Indeed we may assume that $F$ is positive, increasing and concave on some convex open set $\Gamma \subset \mathbb{S}^n$, which is closed under addition of the positive cone. Here $\mathbb{S}^n$ denotes the subspace of $\mathbb{R}^n \times \mathbb{R}^n$ consisting of the symmetric matrices. Suppose also that

$$\text{trace } F_r \to \infty \quad \text{as } \lambda_{\max}(r) \to \infty \tag{2.1}$$

on subsets of $\Gamma$ where $F \geq \delta$ for any $\delta > 0$. The following estimate extends Theorem 4.1 and Remark 4.1 in [23].

**Theorem 2.1.** *Let $u \in C^4(\Omega)$, $D^2u + A \in \Gamma$, be a solution of* (1.1)*, with $A$ strictly regular on the set $\mathcal{U} = \{(x, z, p) \mid x \in \Omega, z = u(x), p = Du(x)\}$. Then for any $\Omega' \subset\subset \Omega$ we have*

$$\sup_{\Omega'} |D^2u| \leq C, \tag{2.2}$$

*where $C$ is a constant depending on $\Omega'$, $\Omega$, $A$, $B$ and $|u|_{1;\Omega}$.*

**2.2. Dirichlet problem.** We present a global second derivative estimate for solutions of the Dirichlet problem, or first boundary value problem, for the Monge–Ampère type equation (1.4). First we introduce a convexity condition for domains, which was fundamental in our applications to optimal transportation in [23], [31]. Namely if $\Omega$ is a connected domain in $\mathbb{R}^n$, with $\partial\Omega \in C^2$, and $A \in C^1(\Omega \times \mathbb{R} \times \mathbb{R}^n; \mathbb{S}^n)$ we say that $\Omega$ is *$A$-convex* (*uniformly $A$-convex*), with respect to $\mathcal{U} \subset \Omega \times \mathbb{R} \times \mathbb{R}^n$, if

$$\left[D_i\, \gamma_j(x) + A_{ij,p_k}(x, z, p)\, \gamma_k(x)\right] \tau_i\, \tau_j \geq 0, \quad (\delta_0), \tag{2.3}$$

for all $x \in \partial\Omega$, $x, z, p \in \overline{\mathcal{U}}$, unit outer normal $\gamma$ and unit tangent vector $\tau$ (for some $\delta_0 > 0$).

When $A \equiv 0$, $A$-convexity reduces to the usual convexity. We also say that a domain $\Omega$ is *$A$-bounded*, with respect to $\mathcal{U} \subset \Omega \times \mathbb{R} \times \mathbb{R}^n$, if there exists a function $\varphi \in C^2(\Omega)$ satisfying

$$D^2\varphi(x) + A_{p_k}(x, z, p)\, D_k\, \varphi \geq \delta_0\, I \tag{2.4}$$

for all $x \in \Omega$, $(x, z, p) \in \mathcal{U}$. Note that when $A \equiv 0$, any bounded domain is $A$-bounded. A domain $\Omega$ is then both uniformly $A$-convex and $A$-bounded if there exists a defining function $\varphi \in C^2(\overline{\Omega})$, satisfying $\varphi = 0$ on $\partial\Omega$, $D\varphi \neq 0$ on $\partial\Omega$, together with (2.4).

**Theorem 2.2.** *Let $u \in C^4(\overline{\Omega})$ be an elliptic solution of equation* (1.4) *in $\Omega$, satisfying $u = g$ on $\partial\Omega$, where $\partial\Omega \in C^4$, $g \in C^4(\overline{\Omega})$. Suppose that $A$ is regular on the set $\mathcal{U} = \{(x, z, p) \mid x \in \Omega, z = u(x), p = Du(x)\}$, with $\Omega$ uniformly $A$-convex and $A$-bounded with respect to $\mathcal{U}$. Then we have the estimate,*

$$\sup_{\Omega} |D^2u| \leq C, \tag{2.5}$$

*where $C$ is a constant depending on $A$, $B$, $\Omega$, $\varphi$ and $|u|_{1;\Omega}$.*

**2.3. Second boundary value problem.** Now we turn our attention to the prescribed Jacobian equation, in the form (1.9). The second boundary value problem, or natural boundary condition, involves the prescription of the image of the mapping $T_u$ in (1.6), that is

$$T(\Omega) = \Omega^* \tag{2.6}$$

for some given domain $\Omega^* \subset \mathbb{R}^n$. If the positive function $\psi$ is given by

$$\psi(x, z, p) = f(x)/g \circ Y(x, z, p) \tag{2.7}$$

for positive $f, g \in C^0(\Omega), C^0(\Omega^*)$ respectively, and $T$ is a diffeomorphism (for example when $\Omega$ is convex), we obtain the necessary condition for solvability,

$$\int_\Omega f = \int_{\Omega^*} g, \tag{2.8}$$

which is the *mass balance condition* in optimal transportation. Following our previous notions of domain convexity, we will say that $\Omega$ is *$Y$-convex* (*uniformly $Y$-convex*, *$Y$-bounded*) with respect to $\Omega^*$ if $\Omega$ is $A$-convex (uniformly $A$-convex, $A$-bounded) with respect to $\mathcal{U}_Y = \{(x, z, p) \mid x \in \Omega, Y(x, z, p) \in \Omega^*\}$, where the matrix function $A$ is given by

$$A = Y_p^{-1}(Y_x + Y_z \otimes p) \tag{2.9}$$

as in equation (1.9). The target domain $\Omega^*$ is *$Y^*$-convex*, with respect to $\Omega$, if for each $(x, z) \in \Omega \times \mathbb{R}$, the set

$$\mathcal{P}(x, z) = \{p \in \mathbb{R}^n \mid Y(x, z, p) \in \Omega^*\} \tag{2.10}$$

is *convex* in $\mathbb{R}^n$ and *uniformly $Y^*$-convex*, with respect to $\Omega$, if $\mathcal{P}(x, z)$ is uniformly convex, with respect to $(x, z) \in \Omega \times \mathbb{R}$. Note that for $\partial\Omega^* \in C^2$, these concepts may also be expressed in the form (2.3), and that when $Y$ is generated by a cost function, which happens when $Y_z \equiv 0$, by virtue of the assumed symmetry of $A$, they are dual to each other (see Section 3).

**Theorem 2.3.** *Let $u \in C^4(\overline{\Omega})$ be an elliptic solution of equation* (1.4) *in $\Omega$, satisfying* (2.6)*, where $\partial\Omega, \partial\Omega^* \in C^4$ and $\Omega, \Omega^*$ are uniformly $Y$-convex, $Y^*$-convex with respect to each other. Suppose also that $\Omega$ is $Y$-bounded and that $A$ is regular on $\mathcal{U}_Y$. Then we have the estimate*

$$\sup_\Omega |D^2 u| \leq C, \tag{2.11}$$

*where $C$ depends on $Y$, $\psi$, $\Omega$, $\Omega^*$ and $|u|_{1;\Omega}$.*

### 2.4. Remarks

**1.** Estimates in $C^3(\overline{\Omega})$ automatically follow from the assumed data regularity in Theorems 2.2 and 2.3, by virtue of the global $C^{2,\alpha}$ estimates [18] and [21]. Classical existence theorems then follow by the method of continuity under additional hypotheses to control the solutions and their gradients.

**2.** The boundary condition (2.6) is a nonlinear *oblique* boundary condition of the form

$$G(x, u, Du) := \varphi^* \circ Y(x, u, Du) = 0, \tag{2.12}$$

where $\varphi^*$ is a defining function for $\Omega^*$. If $|\nabla\varphi^*| = 1$ on $\partial\Omega^*$ we obtain, for $c^{i,j} = D_{p_i} Y^j$,

$$\chi := \gamma \, . G_p(x, u, Du) = c^{i,j} \, \gamma_i \, \gamma_j^* > 0, \tag{2.13}$$

by virtue of ellipticity, and the geometric conditions on $\Omega$ and $\Omega^*$ are used to estimate $\chi$ from below, [31].

**3.** The special cases $A \equiv 0$ of the standard Monge–Ampère equation in Theorems 2.2 and 2.3 are due to Ivochkina [17], Krylov [18], Caffarelli, Nirenberg and Spruck [5] (Theorem 2.2), and Caffarelli [2] and Urbas [32] (Theorem 2.3). Sharp versions for Hölder continuous inhomogeneous terms were proved by Trudinger and Wang [29] and Caffarelli [2].

**4.** Theorems 2.1, 2.2 and 2.3 extend to non-symmetric matrices $A$ in two dimensions.

**5.** The condition of uniform $A$-convexity in Theorem 2.2 may be replaced by the more general condition that there exists a strict sub-solution taking the same boundary conditions, as for the case $A \equiv 0$ in [12].

## 3. Optimal transportation

Let $\Omega$ and $\Omega^*$ be bounded domains in $\mathbb{R}^n$ and $f, g$ nonnegative functions in $L^1(\Omega)$, $L^1(\Omega^*)$ respectively satisfying the mass balance condition (2.8). Let $c \in C^0(\mathbb{R}^n \times \mathbb{R}^n)$ be a cost function. The corresponding Monge–Kantorovich problem of optimal transportation is to find a measure preserving mapping $T_0$ which maximizes (or minimizes) the cost functional,

$$\mathcal{C}(T) = \int_\Omega f(x) \, c(x, T(x)) \, dx, \tag{3.1}$$

over the set $\mathcal{T}$ of measure preserving mappings $T$ from $\Omega$ to $\Omega^*$. A mapping T is called measure preserving if it is Borel measurable and for any Borel set $E \subset \Omega^*$,

$$\int_{T^{-1}(E)} f = \int_E g. \tag{3.2}$$

For the basic theory the reader is referred to the accounts in works such as [9], [24], [33], [34]. To fit the exposition in our previous sections, we consider maximization problems rather than minimization, noting that it is trivial to pass between them replacing $c$ by $-c$.

**3.1. Kantorovich potentials.** The dual functional of Kantorovich is defined by

$$I(u, v) = \int_\Omega f(x) \, u(x) \, dx + \int_{\Omega^*} g(y) \, v(y) \, dy, \tag{3.3}$$

for $(u, v) \in K$ where

$$K = \{(u, v) \mid u \in C^0(\Omega), v \in C^0(\Omega^*), \\ u(x) + v(y) \geq c(x, y), \text{ for all } x \in \Omega, y \in \Omega^*\}. \tag{3.4}$$

It is readily shown that $\mathcal{C}(T) \leq I(u, v)$, for all $T \in \mathcal{T}$, $u, v \in K$. To solve the Monge–Kantorovich problem, we assume $c \in C^2(\mathbb{R}^n \times \mathbb{R}^n)$ and that for each $x \in \Omega$, $p \in \mathbb{R}^n$ there exists a unique $y = Y(x, p)$ satisfying (1.10), together with the corresponding condition for $x$ replaced by $y \in \Omega^*$ and

$$|\det c_{x,y}| \geq c_0 \tag{3.5}$$

on $\Omega \times \Omega^*$ for some constant $c_0 > 0$. Then there exist semi-convex functions $(u, v) \in K$ and a mapping $T = T_u$, given by

$$T_u = Y(\cdot, Du) \tag{3.6}$$

almost everywhere in $\Omega$, such that

$$\mathcal{C}(T) = I(u, v). \tag{3.7}$$

The functions $u, v$, which are uniquely determined up to additive constants, are called *potentials* and are related by

$$u(x) = \sup_{\Omega^*}\{c(x, \cdot) - v(\cdot)\}, \quad v(y) = \sup_{\Omega}\{c(\cdot, y) - u(\cdot)\}. \tag{3.8}$$

Furthermore, for positive densities $f$, $g$, the potential function $u$ will be an elliptic solution of equation (1.11) almost everywhere in $\Omega$ (at points where it is twice differentiable). If $u \in C^2(\overline{\Omega})$, then $u$ is a classical solution of the second boundary value problem (2.6).

**3.2. Interior regularity.** Consistent with our definitions in Section 2, $\Omega$ is $c$-convex with respect to $\Omega^*$ if $c_y(\cdot, y)(\Omega)$ is convex for all $y \in \Omega^*$ and $\Omega^*$ is $c^*$-convex if $c_x(x, \cdot)(\Omega^*)$ is convex for all $x \in \Omega$. As a consequence of Theorem 2.1, we have the main result in [23].

**Theorem 3.1.** *Let $\Omega, \Omega^*$ be bounded domains in $\mathbb{R}^n$, $f \in C^2(\Omega) \cap L^\infty(\Omega)$, $g \in C^2(\Omega^*) \cap L^\infty(\Omega^*)$, $\inf f, \inf g > 0$. Let $c \in C^4(\mathbb{R}^n \times \mathbb{R}^n)$ and $\Omega^*$ be $c^*$-convex with respect to $\Omega$. Suppose that $A$ is strictly regular on the set $\mathcal{U}_Y$, where $A(x, p) = -D_x^2 c(x, Y(x, p))$ and $Y$ is given by* (1.10)*. Then the potential $u \in C^3(\Omega)$.*

**3.3. Global regularity.** From the global second derivative estimate, Theorem 2.3, we obtain a global regularity result, corresponding to Theorem 3.1, which is proved in [31]. For its formulation we say that $\Omega$ ($\Omega^*$) is *uniformly c-convex* (*c*-convex*), with respect to $\Omega^*$, ($\Omega$), if the images $c_y(\cdot, y)(\Omega)$ ($c_x(x, \cdot)(\Omega^*)$) are uniformly convex with respect to $y \in \Omega^*$ ($x \in \Omega$). This agrees with our previous definitions in terms of the vector field $Y$ and matrix $A$ determined by $c$.

**Theorem 3.2.** *Let $\Omega$ and $\Omega^*$ be bounded $C^4$ domains in $\mathbb{R}^n$, $f \in C^2(\overline{\Omega})$, $g \in C^2(\overline{\Omega^*})$, $\inf f > 0$, $\inf g > 0$. Let $c \in C^4(\mathbb{R}^n \times \mathbb{R}^n)$ and let $\Omega$, $\Omega^*$ be uniformly c-convex, $c^*$-convex with respect to each other. Suppose also that $\Omega$ is $Y$-bounded and $A$ is regular on $\mathcal{U}_Y$. Then the potential function $u \in C^3(\overline{\Omega})$.*

### 3.4. Remarks

**1.** For the case of quadratic cost functions,

$$c(x, y) = x.y, \quad Y(x, p) = p, \quad A \equiv 0, \tag{3.9}$$

Theorem 3.1 is due to Caffarelli [1], Theorem 3.2 is due to Caffarelli [2] and Urbas [32]. Note that this case is excluded from Theorem 3.1 but embraced by Theorem 3.2. The interior estimate (2.2) is not valid when $A \equiv 0$.

**2.** By exploiting the geometric interpretation of strict regularity, Loeper [22] has shown that the potential $u \in C^{1,\alpha}(\Omega)$ for certain $\alpha > 0$, when the smoothness of the densities $f, g$ is dropped. Moreover he has shown that the regularity of $A$ is a necessary condition for $u \in C^1(\Omega)$ for arbitrary smooth positive densities.

**3.** As shown in [23], the $c^*$-convexity of $\Omega^*$ is also necessary for interior regularity for arbitrary smooth positive densities.

**4.** The condition of $Y$-boundedness may be dropped in Theorem 2.3 in the optimal transportation case [31].

**5.** Various examples of cost functions for which $A$ is regular or strictly regular are presented in [23] and [31].

## 4. Conformal geometry

In recent years the Yamabe problem for the $k$-curvature of the Schouten tensor, or simply the $k$-Yamabe problem, has been extensively studied. Let $(\mathcal{M}, g_0)$ be a smooth compact manifold of dimension $n > 2$ and denote by Ric and $R$ respectively the Ricci tensor and scalar curvature. The $k$-Yamabe problem is to prove the existence of a conformal metric $g = g_u = e^{-2u} g_0$ such that

$$\sigma_k(\lambda(S_g)) = 1 \quad \text{on } \mathcal{M}, \tag{4.1}$$

where $k = 1, \ldots, n$, $\lambda = (\lambda_1, \ldots, \lambda_n)$ denotes the eigenvalues of $S_g$ with respect to the metric $g$, $\sigma_k$ is the $k$th elementary symmetric function given by

$$\sigma_k(\lambda) = \sum_{i_1 < \cdots < i_k} \lambda_{i_1} \cdots \lambda_{i_k}, \tag{4.2}$$

and $S_g$ is the Schouten tensor of $(\mathcal{M}, g)$ given by

$$\begin{aligned} S_g &= \frac{1}{n-2}\Big(\operatorname{Ric} g - \frac{R_g}{2(n-1)}g\Big) \\ &= \nabla^2 u + \nabla u \otimes \nabla u - \tfrac{1}{2}|\nabla u|^2 g_0 + S_{g_0}. \end{aligned} \tag{4.3}$$

Accordingly we obtain the equation

$$\mathcal{F}_k[u] = F_k^{1/k}\big\{g_0^{-1}\left(\nabla^2 u + \nabla u \otimes \nabla u - \tfrac{1}{2}|\nabla u|^2 g_0 + S_{g_0}\right)\big\} = e^{-2u}, \tag{4.4}$$

where $F_k(r)$ denotes the sum of the $k \times k$ principal minors of the matrix $r \in \mathbb{S}^n$, which is elliptic for $\lambda(S_g) \in \Gamma_k$ where $\Gamma_k$ is the cone in $\mathbb{R}^n$ given by

$$\Gamma_k = \{\lambda \in \mathbb{R}^n \mid \sigma_j(\lambda) > 0,\ j = 1, \ldots, k\}, \tag{4.5}$$

(see for example [3], [28]). When $k = 1$, we arrive at the well-known Yamabe problem [27], that was completely resolved by Schoen in [25]. Note that for Euclidean space $\mathbb{R}^n$, we have

$$S_g = D^2 u + Du \otimes Du - \tfrac{1}{2}|Du|^2 I, \tag{4.6}$$

in agreement with (1.13).

The operators $\mathcal{F}_k$ are *strictly regular* in the sense of (1.18) so that interior estimates corresponding to Theorem 2.1 are readily proven, [35], [13], [7], [14]. However crucial ingredients in the solution of the $k$-Yamabe problem are estimates in terms of $\inf u$ only. These were obtained by Guan and Wang [14] and recently simplified by Chen [8], who derived the gradient estimates directly from the second derivative estimates using $\sigma_1(\lambda) \geq 0$. The following theorem, due to Sheng, Trudinger and Wang [26] ($k \leq n/2$), Gursky and Viaclovsky [16] ($k > n/2$) concerns the solvability of the higher order Yamabe problem, $k > 1$.

**Theorem 4.1.** *Let $(\mathcal{M}, g_0)$ be a smooth compact manifold of dimension $n > 2$ and suppose there exists some metric $g_1$ conformal to $g_0$ for which $\lambda(S_{g_1}) \in \Gamma_k$. Then there exists a conformal metric $g$ satisfying* (4.1) *if either $k > n/2$ or $k \leq n/2$ and* (4.1) *has variational structure, that is it is equivalent to an Euler equation of a functional.*

We remark that (4.1) is variational for $k = 1, 2$ and if $(\mathcal{M}, g_0)$ is locally conformally flat otherwise. The case of Theorem 4.1 when $k = 2$, $n = 4$ was proved in the pioneering work of Chang, Yang and Gursky [6], while the locally conformally flat case was proved by Guan and Wang [15] and Li and Li [19], [20]. The cases $k = 2$, $n > 8$ were obtained independently by Ge and Wang [11]. The reader is referred to the various papers cited above for further information. Also a more elaborate treatment of the case $k > n/2$ is presented in [30].

## References


[1] Caffarelli, L., The regularity of mappings with a convex potential. *J. Amer. Math. Soc.* **5** (1992), 99–104.

[2] Caffarelli, L., Boundary regularity of maps with convex potentials II. *Ann. of Math.* **144** (3) (1996), 453–496.

[3] Caffarelli, L., Nirenberg, L., Spruck, J., The Dirichlet problem for nonlinear second order elliptic equations I. Monge-Ampère equation. *Comm. Pure Appl. Math.* **37** (1984), 369–402.

[4] Delanoë, Ph., Classical solvability in dimension two of the second boundary value problem associated with the Monge-Ampère operator. *Ann. Inst. Henri Poincaré, Analyse Non Linéaire* **8** (1991), 443–457.

[5] Caffarelli, L. A., Nirenberg, L., Spruck, J., Dirichlet problem for nonlinear second order elliptic equations III. Functions of the eigenvalues of the Hessian. *Acta Math.* **155** (1985), 261–301.

[6] Chang, A., Gursky, M., Yang, P., An equation of Monge-Ampère type in conformal geometry, and four-manifolds of positive Ricci curvature. *Ann. of Math.* (2) **155** (2002), 709–787.

[7] Chang, A., Gursky, M., Yang, P., An a priori estimate for a fully nonlinear equation on four-manifolds. *J. Anal. Math.* **87** (2002), 151–186.

[8] Chen, S. S., Local estimates for some fully nonlinear elliptic equations. Preprint; http://arxiv.org/abs/math.AP/0510652.

[9] Gangbo, W., McCann, R. J., The geometry of optimal transportation. *Acta Math.* **177** (1996), 113–161.

[10] Gilbarg, D., Trudinger, N. S., *Elliptic partial differential equations of second order*. Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin 1983.

[11] Ge, Y., Wang, G., On a fully nonlinear Yamabe problem. *Ann. Sci. École Norm. Sup.* (3), to appear.

[12] Guan, B., Spruck, J., Boundary value problems on $S^n$ for surfaces of constant Gauss curvature. *Ann. of Math.* **138** (1993), 601–624

[13] Guan, P., Wang, X. J., On a Monge-Ampère equation arising in geometric optics. *J. Differential Geom.* **48** (1998), 205–223.

[14] Guan, P., Wang, G., Local estimates for a class of fully nonlinear equations arising from conformal geometry. *Internat. Math. Res. Notices* **2003** (26) (2003), 1413–1432.

[15] Guan, P., Wang, G., A fully nonlinear conformal flow on locally conformally flat manifolds. *J. Reine Angew. Math.* **557** (2003), 219–238.

[16] Gursky, M., Viaclovsky, J., Prescribing symmetric functions of the eigenvalues of the Ricci tensor. *Ann. of Math.*, to appear.

[17] Ivochkina, N., A priori estimate of $||u||_{C^2(\overline{\Omega})}$ of convex solutions of the Dirichlet problem for the Monge-Ampère equation. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov* (*LOMI*) **96** (1980), 69–79; English. transl. *J. Soviet Math.* **21** (1983), 689–697.

[18] Krylov, N. V., *Nonlinear elliptic and parabolic equations of the second order*. Math. Appl. (Soviet Ser.) 7, Reidel, Dordrecht, Boston 1987.

[19] Li, A., Li, Y. Y., On some conformally invariant fully nonlinear equations. *Comm. Pure Appl. Math.* **56** (2003), 1416–1464.

[20] Li, A., Li, Y. Y., On some conformally invariant fully nonlinear equations II, Liouville, Harnack, and Yamabe. *Acta Math.* **195** (2005), 117–154.

[21] Lieberman, G. M., Trudinger, N. S., Nonlinear oblique boundary value problems for nonlinear elliptic equations. *Trans. Amer. Math. Soc.* **295** (1986), 509–546.

[22] Loeper, G., Continuity of maps solutions of optimal transportation problems. Preprint; http://arxiv.org/abs/math.AP/0504137.

[23] Ma, X. N., Trudinger, N. S., Wang, X.-J., Regularity of potential functions of the optimal transportation problem. *Arch. Rat. Mech. Anal.* **177** (2005), 151–183.

[24] Rachev, S. T., Rüschendorff, L., *Mass transportation problems*. Vol. I, II, Probab. Appl. (N.Y.), Springer-Verlag, New York 1998.

[25] Schoen, R., Conformal deformation of a Riemannian metric to constant scalar curvature. *J. Differential Geom.* **20** (1984), 479–495.

[26] Sheng, W., Trudinger, N. S., Wang, X.-J., The Yamabe problem for higher order curvatures. Preprint; http://arxiv.org/abs/math.DG/0505463.

[27] Trudinger, N. S., Remarks concerning the conformal deformation of Riemannian structures on compact manifolds. *Ann. Scuola Norm. Sup. Pisa* (3) **22** (1968), 265–274.

[28] Trudinger, N. S., On the Dirichlet problem for Hessian equations. *Acta Math.* **175** (1995), 151–164.

[29] Trudinger, N. S., Wang, X.-J., Boundary regularity for the Monge-Ampère and affine maximal surface equations. *Ann. of Math.*, to appear.

[30] Trudinger, N. S., Wang, X.-J., On Harnack inequalities and singularities of admissible metrics in the Yamabe problem. Preprint; http://arxiv.org/abs/math.DG/0509341.

[31] Trudinger, N. S., Wang, X.-J., On the second boundary value problem for Monge-Ampère type equations and optimal transportation. Preprint; http://arxiv.org/abs/math.AP/0601086.

[32] Urbas, J., On the second boundary value problem for equations of Monge-Ampère type. *J. Reine Angew. Math.* **487** (1997), 115–124.

[33] Urbas, J., *Mass transfer problems*. Lecture Notes, Universität Bonn, 1998.

[34] Villani, C., *Topics in optimal transportation*. Grad. Stud. Math. 58, Amer. Math. Soc., Providence, RI, 2003.

[35] Wang, X. J., On the design of a reflector antenna. *Inverse Problems* **12** (1996), 351–375.

[36] Wang, X. J., On the design of a reflector antenna II. *Calc. Var. Partial Differential Equations* **20** (2004), 329–341.

Neil S. Trudinger, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

E-mail: Neil.Trudinger@anu.edu.au

# The initial value problem for nonlinear Schrödinger equations

Luis Vega*

**Abstract.** I will review some recent work done in collaboration with C. E. Kenig, G. Ponce and C. Rolvung on a general method to solve locally in time the initial value problem for non-linear Schrödinger equations under some natural hypotheses of decay and regularity of the coefficients. Also some non-trapping conditions of the solutions of the hamiltonian flow associated to the initial data is needed. We will not assume ellipticity on the matrix of the leading order coefficients but just a non-degeneracy condition. The method is based on energy estimates which can be performed thanks to the construction of an integrating factor. This construction is of independent interest and relies on the analysis of some new pseudo-differential operators.

**Mathematics Subject Classification (2000).** Primary 35Q20; Secondary 35B45.

**Keywords.** Non-linear Schrödinger equations, ultrahyperbolic operators.

## 1. Introduction

In this lecture I shall describe some joint work with C. E. Kenig and G. Ponce on general non-linear Schrödinger equations built on spatial operators which are given by just non-degenerate quadratic forms. It has been and still is a great pleasure and a privilege to work with both of them.

This research was started some time ago in [14], and we could say it has come to a natural end with [16], [22] and [23], these two latter works written in collaboration with C. Rolvung. In the process we have used some fundamental work done by other authors, in particular those by Hayashi and Ozawa [10], Doi [6], [7], [8], and Craig, Kappeler and Strauss [4]. A look at the introduction of the papers [16], [22], and [23] is enough to realize that the precise results are rather technical and lengthy to write. Therefore, this lecture will be mainly expository, and I refer to the reader to the papers mentioned above and to the review given in [21] for the precise statement of the theorems.

We are interested in solving the initial value problem

$$\begin{cases} \partial_t u = i\tilde{\mathcal{L}}u + F(u, \nabla u, \bar{u}, \nabla \bar{u}), \\ u(x,0) = u_0(x), \end{cases} \tag{1}$$

*This work is partially supported by the grant MTM2004-03029.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

where

- $u(x,t) \in \mathbb{C}$, $x = (x_1, x_2) \in \mathbb{R}^{n_1} \times R^{n_2} = \mathbb{R}^n$, $n = n_1 + n_2$, and $t \in [0, T]$;
- $\tilde{\mathcal{L}}u(x) = \partial_j(\tilde{a}_{jk}(x,t,u,\nabla u,\bar{u},\nabla\bar{u})\partial_k u)$, and $\tilde{A}(u) = (\tilde{a}_{jk})_{jk}$ is a real, symmetric, invertible matrix such that

$$\lim_{|x|\to\infty} \tilde{A} = \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix}; \tag{2}$$

- $F$ is a regular non-linear function, as is a polynomial with no linear terms.

By solving this equation I mean to find a large enough space $X$ of initial data, and a unique solution $u$ for each $u_0 \in X$ up to a time $T = T(u_0)$ such that $u$ is unique in some space $Y$, and the map $u_0 \mapsto u$ from $X$ to $Y$ is continuous. Typically

$$X = X_{\alpha_0\beta_0} = \{u_0 \text{ such that } x^\alpha \partial^\beta u_0 \in L^2;\ |\alpha| \le \alpha_0,\ |\beta| \le \beta_0\},$$
$$Y = Y_{\alpha_1\beta_1} = \{\mathcal{C}([0,T] : X_{\alpha_1\beta_1})\}$$

for some finite $(\alpha_j, \beta_j)$, $j = 0, 1$, and with possibly $0 < \alpha_1 < \alpha_0$, $0 < \beta_1 < \beta_0$.

There are several reasons which motivate the study of such a general initial value problem. First of all, within this general model there are relevant equations which appear in the physics literature, as Davey–Stewartson and Zakharov–Schulman systems, Landau–Lipschitz equations, the Schrödinger map and others; see [32]. These physical models usually have a very rich algebraic structure which in many cases allows for some short cuts in their analysis; in particular regarding existence results. However, uniqueness generally involves considering the PDE solved by the difference of two solutions which need not have the same algebraic structure as the starting one, but that still is under the general setting given in (1). I will be more precise about this point in Section 3, where the particular case of the Schrödinger map is analyzed.

Another motivation is to get a better understanding of ultrahyperbolic operators including the constant coefficient case

$$\mathcal{L}_0 = \Delta_{x_1} - \Delta_{x_2}.$$

Notice that neither the heat nor the wave flow can be defined for $\mathcal{L}_0$, while the Schrödinger flow

$$e^{it\mathcal{L}_0} \tag{3}$$

makes perfect sense. Our knowledge about this operator is far behind the classical one $e^{it\Delta}$. In Section 9, I gather some information and open questions about linear and non-linear perturbations of the free propagator (3).

The rest of the paper is devoted to explain the assumptions needed to solve (1), and to exhibit the algebraic tools we use to construct the solution. The main result is given in Section 7, and some remarks about the elliptic setting can be found in Section 8.

## 2. Energy estimates

In order to solve equation (1) we use the so-called energy method. This is based on three steps. The first one is to add some artificial viscosity to the right-hand side of (1) depending on $\varepsilon > 0$, and to consider for some $T_\varepsilon > 0$ the equation

$$\begin{cases} \partial_t u_\varepsilon = -\varepsilon \Delta_x^2 + i\tilde{\mathcal{L}} u + F, & t \in [0, T_\varepsilon), \\ u_\varepsilon = u_0, & t = 0. \end{cases} \tag{4}$$

The existence of a solution $u_\varepsilon$ of (4) can be easily proved by Picard iteration, using the regularity properties of the free propagator $e^{-\varepsilon t \Delta^2}$.

The second step relies on proving energy estimates for $u_\varepsilon$ independent of $\varepsilon$. These estimates will give a universal time of existence valid for all $u_\varepsilon$, and also will allow us to pass to the limit in $\varepsilon$ to obtain a solution of (1).

The final step is to prove uniqueness by looking at the equation satisfied by the difference $u_{\varepsilon\varepsilon'} = u_\varepsilon - u_{\varepsilon'}$. It is at this point where to work in a general setting as (1) turns out to be fundamental because the equation for $u_{\varepsilon\varepsilon'}$ is again of the same type.

In the rest of the paper I will mainly focus on the question of the energy estimates, exhibiting the algebraic tools needed to obtain them.

Recall that we have already built the solution $u_\varepsilon$ of (4) with $\tilde{\mathcal{L}}$ as in (1). Hence we can now understand (4) as a linear equation

$$\begin{cases} \partial_t u = \mathcal{L}_\varepsilon u + i\big(\overrightarrow{b_1} \cdot \nabla u + \overrightarrow{b_2} \cdot \nabla \bar{u} + c_1 u + c_2 \bar{u}\big) + f, \\ u(x, 0) = u_0(x), \end{cases} \tag{5}$$

with

$$\mathcal{L}_\varepsilon u = \partial_t a_j(x,t)\partial_k u - \varepsilon \Delta^2 u; \;\; A = \big(a_{jk}(x,t)\big)_{jk} = \big(\tilde{a}_{jk}(x,t,u,\nabla u, \bar{u}, \nabla \bar{u})\partial_k u\big)_{jk}.$$

Therefore it seems appropriate to assume the following hypotheses regarding $A$ and the coefficients $\overrightarrow{b_1}$, $\overrightarrow{b_2}$ and $c_1, c_2$:

H1. $A$ is a regular real symmetric non-degenerate matrix, i.e. there exists $\gamma_0 \in (0, 1)$ such that

$$\gamma_0 |\xi| \leq |A\xi| \leq \gamma_0^{-1} |\xi|.$$

H2. The coefficients $\partial_x a_{jk}$, $\partial_{tx} a_{jk}$, $\partial_t a_{jk}$ have a pointwise decay for a sufficiently large $|x|$,

$$\sup_{|t| \leq T} |\partial_x^\alpha a_{jk}(x,t)| + |\partial_t \partial_x^{\alpha'} a_{jk}(x,t)| \leq \frac{C}{(1+|x|)^N},$$

with $N$, $|\alpha|$ and $|\alpha'|$ large enough, and so that

$$A - \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix}$$

has a similar decay for a large $|x|$.

H3. $\overrightarrow{b_1}$ and $\overrightarrow{b_2}$ are smooth complex vector fields that decay pointwise together with their derivatives at infinity.

H4. $c_1$ and $c_2$ are smooth complex scalar fields bounded together with their derivatives at infinity.

Let us start assuming that in (5) the external forces $f$ and the potentials $\overrightarrow{b_1}$, $\overrightarrow{b_2}$ and $c_1$, $c_2$ are zero. Then because $A$ is real and symmetric we trivially have

$$\begin{aligned}\frac{d}{dt}\langle u, u\rangle &= \langle \mathcal{L}_\varepsilon u, u\rangle + \langle u, \mathcal{L}_\varepsilon u\rangle \\ &= -\langle iA\partial_x u, \partial_x u\rangle - \langle \partial_x u, iA\partial_x u\rangle - 2\varepsilon\langle \Delta u, \Delta u\rangle \\ &= -2\varepsilon\langle \Delta u, \Delta u\rangle \le 0.\end{aligned} \tag{6}$$

Therefore

$$\sup_{0\le t\le T}\langle u, u\rangle + 2\varepsilon\int_0^T \langle \Delta u, \Delta u\rangle dt \le \langle u(0), u(0)\rangle. \tag{7}$$

If in the above calculation we add zero order terms $c_1$ and $c_2$ we will obtain after integration in time

$$\langle u, u\rangle(t) \le e^{Mt}\langle u, u\rangle(0), \tag{8}$$

with $M = \sup(|c_1| + |c_2|)$. As a conclusion from the point of view of energy estimates zero order terms are harmless.

Let us know consider that first order terms are not trivial. It is straightforward that if $\operatorname{Re}\overrightarrow{b_1}$ is zero and $\overrightarrow{b_2}$ just bounded, an integration by parts as the one given in (6) leads to the same estimate. However, the situation is completely different if $\operatorname{Re}\overrightarrow{b_1} \neq 0$. This can be easily seen even in one dimension.

Consider the model problem

$$\partial_t u = i(\partial_x^2 u + b_1\partial_x u) - \varepsilon\partial_x^4 u, \quad x \in \mathbb{R}. \tag{9}$$

Take for simplicity $b_1 = 1$. Using $\hat{u}$ the Fourier transform of $u$ and Parseval's identity the calculation given in (6) becomes

$$\begin{aligned}\frac{d}{dt}\langle u, u\rangle &= \langle (i\partial_x^2 - \varepsilon\partial_4^x + i\partial_x)u, u\rangle + \langle u, i\partial_x^2 - \varepsilon\partial_x^4 + i\partial_x\rangle \\ &= -\varepsilon\langle \xi^4\hat{u}, \hat{u}\rangle - \langle \xi\hat{u}, \hat{u}\rangle.\end{aligned} \tag{10}$$

Therefore if $\operatorname{supp}\hat{u} \subset (-\infty, 0)$ we can not obtain a uniform bound in $\varepsilon$ for $\langle u, u\rangle$. This is usually called the loss of derivatives obstruction, because regardless of $\varepsilon$ there is in (10) one derivative more on the right-hand side than on the left-hand side. The final conclusion is that we should get rid, if possible, of $\operatorname{Re} b_1$.

Nevertheless there is a simple way of removing the first order term in (9) creating other ones of order zero, which is to use the integrating factor $\exp\left(\frac{1}{2}\int_x^\infty b_1\right)$. For this purpose define

$$v(x, t) = \mathcal{K}u = e^{1/2\int_x^\infty b_1(y,t)dy}u(x, t). \tag{11}$$

The equation for $v$ becomes

$$\partial_t v = i\partial_x^2 v - \varepsilon \mathcal{K} \partial_x^4 u + c(x,t)v, \tag{12}$$

with

$$c(x,t) = \frac{i}{2}\partial_x b_1 + i\frac{b_1^2}{4} + \frac{i}{2}\int_x^\infty \partial_t b_1(y,t)dy. \tag{13}$$

Notice that $\mathcal{K}$ given in (11) is invertible so that

$$\varepsilon \mathcal{K}\partial_x^4 u = \mathcal{K}\partial_x^4 \mathcal{K}^{-1} v = \varepsilon \partial_x^4 v + \varepsilon \text{ (lower order)}.$$

This lower order terms are harmless because they are of order $\varepsilon$ and can be easily absorbed using (7) and taking $T$ small enough. Looking carefully at the calculation above we see that the only assumption we need on $b_1$ is that

$$\sup_{x,\, 0\le t\le T} \left|\operatorname{Re}\int_x^\infty b_1(y,t)\right| + \left|\operatorname{Re}\int_x^\infty \partial_t b_1(y,t)dt\right| \le M. \tag{14}$$

Notice also that because the operator $\mathcal{K}$ is given by multiplication by the integrating factor we have

$$\mathcal{K}\bar{u} = \overline{\mathcal{K}u}. \tag{15}$$

Therefore the symmetry of the good terms $\operatorname{Im} b_1$ and $b_2$ is not destroyed and can be added to (9) so that the computation given above works the same without any extra difficulty.

Finally let us recall that the condition given in (14) is very reminiscent of the one obtained by Mizohata in [27] (see also [11]), and suggests that H3 and H4 are natural assumptions.

## 3. The Schrödinger map

The above analysis rises the natural question of whether or not the integrating factor can be constructed also in higher dimensions. It is at this stage where considering a specific equation and not the general case given in (1) can make a big difference. An illustrative example is the so-called Schrödinger map given by

$$\begin{cases} \overrightarrow{u_t} = \overrightarrow{u} \wedge \Delta \overrightarrow{u}, \quad \overrightarrow{u} = \overrightarrow{u}(x,t), \quad x \in \mathbb{R}^d,\ t \in \mathbb{R},\ d = 1,2, \\ |\overrightarrow{u}|^2 = 1. \end{cases} \tag{16}$$

This equation written in coordinates (for example using the stereographic projection) involves non-linearities in the first order terms. Let us try to understand first the

one dimensional case $d = 1$. Then $\overrightarrow{u}$ is nothing but the tangent vector to a three dimensional curve $\overrightarrow{\gamma}(x,t)$ (i.e. $\overrightarrow{u} = \overrightarrow{\gamma}_x$) which satisfies

$$\begin{cases} \overrightarrow{\gamma}_t = \overrightarrow{\gamma}_x \wedge \overrightarrow{\gamma}_{xx}, \quad x \in \mathbb{R},\ t \in \mathbb{R}, \\ |\overrightarrow{\gamma}_x|^2 = 1. \end{cases} \tag{17}$$

This equation, which is sometimes called the Localized Induction Approximation, was obtained for the first time by Da Rios in 1906, see [5], as a crude approximation of the evolution of a vortex filament within Euler equations. From a geometrical point of view is better to call it the binormal flow because using Frenet equations (17) can be written as

$$\overrightarrow{\gamma}_t = c\,\overrightarrow{b},$$

where $c$ denotes the curvature and $\overrightarrow{b}$ the binormal vector.

In 1972 Hasimoto [9] proposed the transformation

$$\psi(x,t) = c(x,t)e^{i\int_0^x \tau(y,t)\,dy}, \tag{18}$$

with $\tau$ denoting the torsion, to simplify (17) and therefore (16). After some computations he proves that if $\overrightarrow{\gamma}$ satisfies (17) then $\psi$ solves

$$\psi_t = i\psi_{xx} + \frac{i}{2}(|\psi|^2\psi + a(t)), \tag{19}$$

for some real function $a(t)$. Therefore we could understand Hasimoto's transformation (18) as some kind of integrating factor which removes the non-linear first order terms appearing when (16) is written in local coordinates, to the expense of cubic zero order terms which are much easier to handle.

The situation for the Schrödinger map in dimension 2 is much more delicate. Hasimoto's transformation can still be used but it is not possible to remove completely the first order terms; see [1]. The equation obtained after the transformation has a good symmetry from the point of view of energy estimates which allows to prove existence even for $H^1$-solutions. However, much of this symmetry is lost when considering difference of solutions.

Therefore, we are back to the question we started this section with. In the one dimensional case it is possible to construct the integrating factor which greatly simplifies the equations, but a similar transformation even in the two dimensional case is far from clear.

## 4. The integrating factor

Let us go back to our linear equation (5) and to simplify assume in this section that $\varepsilon = 0$. Accordingly we shall write $\mathcal{L}_\varepsilon = \mathcal{L}$. We are looking for an operator $\mathcal{K}$ such that

$$(\mathcal{L}\mathcal{K} - \mathcal{K}\mathcal{L}) = \mathcal{K}\overrightarrow{b_1}\cdot\nabla + \text{ zero order}. \tag{20}$$

It turns out that this equation, except in very simple cases, does not have a solution when just the algebra of classical differential operators is considered, as it happens in the one dimensional case, where $\mathcal{K}$ was constructed using the elemental operations of multiplication and integration. Therefore we need to consider $\mathcal{K}$ to be a pseudo-differential operator $\mathcal{K} = \mathcal{K}(x, t, D)$ given by:

$$\mathcal{K}u = \frac{1}{(2\pi)^n} \int_y \int_\xi k(x, t, \xi) e^{i(x-y)\xi} u d\xi \, dy. \tag{21}$$

If in the expression above we take

$$k(x, t, \xi) = a_\alpha(x, t)(i\xi)^\alpha, \tag{22}$$

then $\mathcal{K} = a_\alpha(x, t)\partial_x^\alpha$. Notice that in this case if $a_\alpha$ is regular and bounded together with its derivatives we get

$$|\partial_x^{\beta_1} \partial_\xi^{\beta_2} k(x, t, \xi)| \leq C_{\beta_1\beta_2}(1 + |\xi|)^{|\alpha|-|\beta_2|}. \tag{23}$$

Assume for a moment that $\mathcal{L}$ is the constant coefficient operator $\mathcal{L}_0$. That is to say

$$A = \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix}.$$

Solving formally in (20) we get that $k$ should be given by

$$k(x, t, \xi) = \exp\left(\frac{1}{2} \int_0^\infty \overrightarrow{b_1}(x + s\tilde{\xi}, t)\right) \cdot \xi ds, \tag{24}$$

with $\tilde{\xi} = A\xi = (\xi_1, -\xi_2)$.

Notice that in (24) $s \mapsto x + s\tilde{\xi}$ is a geodesic associated to the pseudo-metric (or metric if $A = \mathbb{I}_n$) given by $A$. In the variable coefficient case $A = (a_{jk})_{jk}$ we have to define

$$k(x, t, \xi) = \exp\left(\frac{1}{2} \int_0^\infty \overrightarrow{b_1}\,(X(s; x, \xi)) \cdot \Xi(s; x, \xi) ds\right) \tag{25}$$

with $(X(s; x, \xi), \Xi(s; x, \xi))$ solutions of the hamiltonian flow $H_A$ given by

$$\begin{cases} \dfrac{d}{ds} X_j(s; x_0, \xi_0) = -2 \displaystyle\sum_{k=1}^n a_{jk}\,(X(s; x_0, \xi_0) \Xi_k(s; x_0, \xi_0)) \\ \dfrac{d}{ds} \Xi_j\,(s; x_0, \xi_0) = \displaystyle\sum_{k,l=1}^n \partial_j a_{kl}\,(X(s; x_0, \xi_0))\; \Xi_k(s; x_0, \xi_0) \Xi_l(s; x_0, \xi_0) \\ (X(0; x_0, \xi_0), \Xi(0, x_0, \xi_0)) = (x_0, \xi_0). \end{cases} \tag{26}$$

Notice that above we have dropped the dependence on time to simplify the exposition. In fact, in our main result which is given in Section 7, we will impose some hypothesis on $H_A$ where $A$ is determined just by the initial condition $u_0$.

It follows from (26) that

$$\frac{d}{ds}\langle A\Xi(s;x,\xi),\,\Xi\rangle = 0,$$

and therefore in the elliptic case (i.e. $\langle A\xi,\xi\rangle > C|\xi|^2$ with $C>0$) we get that there is a constant $\gamma_0 > 0$ such that

$$\gamma_0|\xi_0|^2 \le |\Xi|^2 \le \gamma_0^{-1}|\xi_0|^2,$$

and $(X(s;x,\xi),\,\Xi(s;x,\xi))$ are globally defined. However, the situation in the non-elliptic case is different and the following properties have to be proved, see [22]:

a) global existence;

b) continuous dependence w.r.t. $(x_0,\xi_0)$ with just a polynomial growth on $|x_0|$;

c) the trajectories are asymptotically free.

In order to obtain these three conditions we will need $A(x,t)$ to verify the hypotheses H1, H2, H3 and H4 given in Section 2 together with the following non-trapping condition.

H5. The solutions $(X,\,\Xi)$ of the hamiltonian flow $H_A$ associated to $A$ verifies that given $M$ and $(x,\xi)\in\mathbb{R}^n\times(\mathbb{R}^n-\{0\})$ there is $s_0$ such that

$$|X(s;x,\zeta)| > M \quad \text{for all } s > s_0.$$

It is easy to justify that H5 is necessary because otherwise the integral given in (24) will not convergence even for compactly supported $\overrightarrow{b_1}$.

This non-trapping condition is not easy to verify, although it is obviously true in the constant coefficient case. Nevertheless it is a stable condition. In [22] the following lemma is proved.

**Lemma 4.1.** *Consider $A(x,t)=A(x)$ such that the hypotheses* H1, H2 *and* H5 *hold. Let $B(x)$ be an $n\times n$ real matrix with entries in the Schwartz class $\mathcal{S}(\mathbb{R}^n)$, and define $A_\delta(x)=A(x)+\delta B(x)$ with $\delta$ so small that there is a constant $\gamma_0>0$ such that for all $\xi\in\mathbb{R}^n$*

$$\frac{\gamma_0}{2}|\xi| \le |A_\delta\xi| \le 2\gamma_0^{-1}|\xi|.$$

*Then there is $\delta_0>0$ such that $A_\delta$ is non-trapping for all $0<\delta<\delta_0$.*

**Remark.** In the statement of the lemma above it is sufficient to assume that the entries of $B$ have a finite number of derivatives with a finite power like decay.

So in the particular case of IVP (1) it will be sufficient to impose the non-trapping condition H5 to

$$A_{u_0}(x) = A(x,0,u_0,\bar{u}_0,\nabla u_0,\nabla\bar{u}_0). \tag{27}$$

## 5. The symbols

Once the hamiltonian flow is built we can construct the "integrating factor" given by the symbol (25). Nevertheless a new problem appears. This symbol is not in any known class of pseudo differential operators even if $A = \mathbb{I}_n$ and $X(x,\xi) = x + s\xi$, $\Xi = \xi$. A model example of a symbol which behaves as the one in (24) and it is easier to handle is, in dimension two,

$$k(x,\xi) = \psi(x \cdot (A\omega)^{\perp})\chi(|\xi|), \quad \omega = \frac{\xi}{|\xi|}, \quad \xi = (\xi_1,\xi_2), \quad \xi^{\perp} = (-\xi_2,\xi_1) \tag{28}$$

with $\psi \in \mathcal{S}(\mathbb{R})$ and $\chi$ regular, $\chi(0) = 0$ and $\chi(t) = 1$ if $t > 1$. We observe that

$$|\partial_\xi^\alpha k| \le \left(\frac{1+|x|}{1+|\xi|}\right)^{|\alpha|}, \tag{29}$$

being the growth in $|x|$ a problem to handle $k(x,\xi)$ (compare this situation with the one exhibited in (23)). In [22] we prove the following result.

**Proposition 5.1.** *Take $\overrightarrow{b_1}$ in the vector valued Schwartz class $\overrightarrow{\mathcal{S}} \in (\mathbb{R}^n)$ and*

$$\begin{aligned} k(x,\xi) &= \exp\left(\frac{1}{2}\int_0^\infty \overrightarrow{b_1}(x + s\tilde{\xi}) \cdot \xi \, ds\right)\chi(|\xi|), \\ \tilde{\xi} &= A\xi, \quad A = \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix}, \end{aligned} \tag{30}$$

*with $\chi(|\xi|)$ as in* (28)*. Then $\mathcal{K}(x,D)$ given in* (21) *is bounded from $L^2$ into $L^2$.*

This proposition and a more general result was proved if $A = \mathbb{I}_n$ (i.e. $n_2 = 0$) by Craig, Kappeler and Strauss [4]. A key part of their argument is the good behaviour of the radial derivatives of the symbol $k$ given in (30). It is easy to check that

$$\left|\left(\frac{\xi}{|\xi|} \cdot \nabla_\xi\right)^\alpha k(x,\xi)\right| \le \frac{C_\alpha}{(1+|\xi|)^\alpha}. \tag{31}$$

This property is still true for general non-degenerate matrices $A$ but it is not sufficient to prove Proposition 5.1. In fact in [4] another geometric assumption on the "essential" support of the symbol $k$ is needed besides (31). This property does not hold when $A$ is not the identity matrix, see [19] for a detailed discussion of this issue.

The $L^2$ estimate given above is not enough. Other results regarding the composition and the computation of the adjoints of these operators have to be proved in order to do the algebraic manipulations we exhibited in the one dimensional case. The results can be seen in [22].

Now it is time to recall that to simplify the exposition we assumed that $\varepsilon = 0$. To avoid this restriction some other properties about the symbols introduced in the above proposition are needed. These were proved in [23]. However the results in [22]

and [23], although sufficient for our purposes, are quite restrictive and the algebraic manipulations we can do with these operators are very rigid.

There is also another important constraint, which is that we are able to handle just symbols as (24) but not the general cases given in (25). The reader could then ask how we overcome this difficulty. The answer is in our previous work done on the I.V.P. (1) in [14] and [15]. That paper deals with non-linearities which are small perturbations of the constant coefficient case. Smallness allows hiding the first order terms thanks to some smoothing properties of the solutions of the corresponding free propagator. These properties are the subject of the next section.

## 6. The local smoothing

In [14] and [15] we proved that if the first order terms in (1) are small they can be handled by the so called local smoothing property of the free operators $e^{it(\Delta_{x_1}-\Delta_{x_2})}$. Notice that this family of operators is reversible in time and leave invariant the Hilbert space $L^2$. Moreover they commute with differentiation, and therefore the classical $L^2$-Sobolev spaces $H^s$ of distributions with $s$ derivatives in $L^2$ also remain invariant under the flow. For this reason there can not be any gain of global derivatives in $L^2$, because otherwise making the flow go backwards we would get a contradiction. It was proved by Kato in [12] and independently by Kruzhkov and Faminskii in [24] that the solutions of the Korteweg–de Vries (KdV) equation

$$\begin{cases} u_t + u_{xxx} + uu_x = 0, \quad u = u(x,t), \quad x \in \mathbb{R},\ t \in \mathbb{R}, \\ u(x,0) = 0 \end{cases}$$

gain for almost every time and locally in space one derivative in $x$ with respect to the initial condition $u_0$. This "local smoothing" is a consequence of the dispersive character of the linear part of the KdV equation, and still holds if the non-linear term $uu_x$ is removed.

It is well known that the free Schrödinger equation is also dispersive, and therefore it should have an analogous smoothing property. In that case the solution gains $1/2$-derivative locally in $x$ and again for a.e. time with respect to the initial data. This property was established independently in [3], [31], and [35], [36], (see also [37]), in the elliptic setting, and in [13] for the general case.

However, this 1/2 gain is not sufficient to deal with first order terms which involve a full derivative. In [14] and [15] it is proved that the solution of

$$\begin{cases} i\partial_t u + \Delta_{x_1} u - \Delta_{x_2} u = F \\ u(x,0) = 0 \end{cases}$$

gains one full derivative with respect to the right-hand side $F$. The proof in [14] strongly uses the Fourier transform and is not adapted to the variable coefficient

situation necessary to treat (1). This was proved later on by S. Doi in [6], [7] and [8]. The results by Doi are remarkably robust, and for example, although he proves it for the elliptic case and for scalar equations, it can be extended without any difficulty to the non-elliptic setting and for systems; see [17], [22] and [23]. The idea is to construct a classical pseudo-differential operator $b(x, D)$ such that the commutator with the general $\mathcal{L} = \partial_j(a_{jk}\partial_k)$ satisfies

$$\begin{aligned}\int_0^T \langle i\,[b(x,D)\mathcal{L} - \mathcal{L}\,b(x,D)]\,u, u\rangle\,dt \\ \geq \frac{c_0}{2}\int_0^T \left\langle \frac{(1-\Delta)^{1/2}u}{1+|x|^2}, u\right\rangle dt - \frac{2}{c_0}T\sup_{0<t<T}\langle u,u\rangle,\end{aligned} \tag{32}$$

for some universal constant $c_0 > 0$. Here I am purposely using "universal" without giving a precise definition. The full argument given in [23] depends on this constant in a crucial way, and I refer to the reader to the introduction of that paper for a more precise statement. It is also important to notice that the above inequality is useful as long as there is a control on the $L^2$ norm given by the term $\langle u, u\rangle$.

## 7. The main result

Our main result in [23] is the following.

**Theorem 7.1.** *Under the hypotheses* H1–H4 *there exists* $N = N(n) \in \mathbb{Z}^+$ *such that given any*

$$u_0 \in H^s(\mathbb{R}^n) \quad \text{with } \langle x\rangle^N \partial_x^\alpha u_0 \in L^2(\mathbb{R}^n),\ |\alpha| \leq s_1, \tag{33}$$

$s, s_1 \in \mathbb{Z}^+$ *sufficiently large, and* $s > s_1 + 4$, *for which the hamiltonian flow* $H_A$ *given in* (26) *associated to the quadratic form*

$$A = A_{u_0}(x,\xi) = \sum_{j.k=1}^n a_{jk}(x, 0, u_0, \bar{u}_0, \nabla u_0, \nabla\bar{u}_0)\xi_j\xi_k \tag{34}$$

*is non-trapping, there exist* $T_0 > 0$, *depending on*

$$\lambda = \|u_0\|_{s,2} + \sum_{|\alpha_1|\leq s_1} \left\|\langle x\rangle^N \partial_x^\alpha u_0\right\|_2$$

*the constants in* H1–H4 *and on the non-trapping condition* H5, *and a unique solution* $u = u(x,t)$ *of the equation* (1) *with initial data* $u(x,0) = u_0(x)$ *on the time interval* $[0, T_0]$ *satisfying*

$$\begin{aligned}&u \in C\big([0,T_0]\,:\,H^{s-1}\big) \cap L^\infty\big([0,T_0]\,:\,H^s\big) \cap C^1\big((0,T_0)\,:\,H^{s-3}\big),\\ &\langle x\rangle^N \partial_x^\alpha u \in C\left([0,T_0]\,:\,L^2\right), \quad |\alpha| \leq s_1.\end{aligned} \tag{35}$$

*Moreover, if* $u_0 \in H^{s'}(\mathbb{R}^n)$ *with* $s' > s$ *then* (35) *holds with* $s'$ *instead of* $s$ *in the same interval* $[0, T_0]$.

We have seen in Section 2 that we can obtain energy estimates as (7) if we are able to get rid of the first order terms, and that we can achieve that using an integrating factor. But also we pointed out at the end of Section 5 that the integrating factor we were able to construct was just for a hamiltonian flow which is free outside of a ball. However as we see in the statement of Theorem 7.1, this is not the case when $H_A$ is given as in (35). The way to bypass this obstruction is to write

$$A = A_R + \left( \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix} - A_R \right)$$

with

$$A_R = \begin{pmatrix} \mathbb{I}_{n_1} & 0 \\ 0 & -\mathbb{I}_{n_2} \end{pmatrix}$$

if $|x| \geq R$ and $R > 0$ is a large parameter to be fixed. The error terms created by this decomposition are of first order and can be done small by taking a large enough $R$. Then the local smoothing inequality (32) can be used as in [14] to control them. This creates the problem of how to handle

$$\sup_{0<t<T} \langle u, u \rangle.$$

But this quantity is precisely the one we started with when doing the energy estimate (7). Notice that in (32) appears multiplied by the factor $T$, and therefore it can be absorbed by the left-hand side of (7) by taking a small enough $T$, which closes the argument.

## 8. The elliptic case

In this section we give some comments that illustrate the substantial differences that appear when in (1) $A$ is the identity matrix. To start with the main result in this case is far more general, and I refer to the reader to the work [17] for a precise statement; see also [29], [28], and [26].

The key difference is the existence of a fundamental argument due to Chihara, see [2], which goes as follows. He first writes (1) as a system in $(u, \bar{u})$. The problem of doing this is that Doi's trick, explained in Section 6, can not be carried out because the pseudo-differential operator $b(x, D)$ needed to prove the estimate (32) does not have the algebraic property (15). Therefore the good structure that the terms

$$\overrightarrow{b_2} \cdot \nabla \bar{u}$$

have for the integration by parts we exhibited in Section 2 is lost after applying the operator $b(x, D)$. Notice these bad terms are off the diagonal. The observation of Chihara is that the corresponding matrix can be easily diagonalized to the expense of zero order terms that as usual are harmless.

In order to explain how this diagonalization is done let us consider the model problem

$$\begin{cases} \partial_t u = i\Delta u + \overrightarrow{b_2}\cdot\nabla\bar{u}. \\ u(x,0) = u_0(x), \end{cases} \tag{36}$$

with $\overrightarrow{b_2} \in \mathbb{C}^n$ a constant vector.

As a system this equation is written as

$$\begin{pmatrix} u \\ \bar{u} \end{pmatrix}_t = \begin{pmatrix} i\Delta & \overrightarrow{b_2}\cdot\nabla \\ \overrightarrow{\bar{b}_2}\cdot\nabla & -i\Delta \end{pmatrix}\begin{pmatrix} u \\ \bar{u} \end{pmatrix}.$$

The eigenvalues of the above matrix are

$$\pm i\big(\Delta^2 - (\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2}.$$

Therefore we are lead to consider the system

$$\begin{pmatrix} v \\ \bar{v} \end{pmatrix}_t = i\begin{pmatrix} \big(\Delta^2 - (\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2} & 0 \\ 0 & -\big(\Delta^2 - (\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2} \end{pmatrix}\begin{pmatrix} v \\ \bar{v} \end{pmatrix}.$$

Notice that

$$\big(\Delta^2 - (\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2} = \Delta\big(1 - \Delta^{-2}(\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2}.$$

The argument ends observing that

$$\Delta\big(1 - \Delta^{-2}(\overrightarrow{b_2}\cdot\nabla)\big(\overrightarrow{\bar{b}_2}\cdot\nabla\big)\big)^{1/2} = \Delta + \text{zero order}. \tag{37}$$

However, the identity (37) is false if the laplacian is changed by an operator of the type $\Delta_{x_1} - \Delta_{x_2}$ and therefore this trick does not work in that situation. Although we have oversimplified the problem considering $\overrightarrow{b_2} \in \mathbb{C}^n$ as a constant vector, the above computations can also be carried out without major difficulty for $\overrightarrow{b_2}(x)$ regular and bounded to the expense of creating zero order terms which as usual are harmless; see [20].

Another important difference of the elliptic setting is that perturbations of the type $\Delta\bar{u}$ are also allowed. The way of seeing this is by a diagonalization argument similar to the one we have just done. Consider $a > 0$ and $b \in \mathbb{C}$ and the equation

$$u_t = ia\Delta u + ib\Delta\bar{u}.$$

Then differentiating with respect to $t$ on both sides we get

$$u_{tt} = (-a^2 + |b|^2)\Delta^2 u$$

which is well posed as long as

$$a^2 > |b|^2.$$

This elemental algebra is much more rigid in the non-elliptic setting and works only for some trivial cases; see [23].

## 9. Remarks on ultrahyperbolic operators

In this final section we gather some information and open questions about linear and non-linear perturbations of the free propagator

$$e^{it\mathcal{L}_0} \quad \text{with } \mathcal{L}_0 = \Delta_{x_1} - \Delta_{x_2}. \tag{38}$$

As we saw in Section 6 one of the fundamental properties of this flow is the local smoothing effect. In fact (32) was a key ingredient to overcome the loss of derivatives obstruction which was explained in Section 2. In order to use (32) one is lead to study the following maximal function

$$\sup_t |e^{it\mathcal{L}_0} u_0|^2, \tag{39}$$

which is defined for all $x \in \mathbb{R}^n$.

In our study of (1) we did not look at the question of which is the minimal regularity to be assumed on the initial condition so that the equation is solvable. This is something which strongly depends on the specific equation one is looking at. Therefore, and for our purposes, the necessary bounds for (39) are rather simple to obtain and there is no difference between the elliptic and the non-elliptic situation at that level. The situation is completely different when looking at a specific model as for example the Schrödinger map I mentioned in Section 3. In that case having sharp bounds for the maximal function can be very useful.

It has been recently proved in [30] that the maximal function given in (39) has a different behaviour for $\mathcal{L}_0$ than for the laplacian, being worse in the former case. Also and with respect to the local smoothing it is known that $\mathcal{L}_0$ is much more sensitive to first order perturbations than the laplacian. In [20] it is proved that the $\frac{1}{2}$ derivative gain of classical Schrödinger flows I mentioned in Section 6 can be reduced to just $\frac{1}{4}$ for $\mathcal{L}_0$.

Regarding non-linear perturbations very little is known about ill-posedness results. In fact and to the best of my knowledge the only one obtained in that direction is about the semilinear equation

$$\begin{cases} \frac{1}{i}\partial_t u = \Delta_{x_1} u - \Delta_{x_2} u \pm |u|^p u \\ u(x,0) = u_0, \end{cases} \tag{40}$$

and is given in [18] with $u_0 = c\delta$. There it is proved that (40) is ill-posed if $p \geq \frac{2}{n}$ (the proof is done for $\Delta$ but it works the same for $\mathcal{L}_0$).

The question of well-posedness of (40) is related to the existence of Strichartz estimates for the free propagator $e^{it\mathcal{L}_0}u_0$ with $u_0 \in L^2$. In that case it is well known that there is no difference between a general $\mathcal{L}_0$ and the laplacian. This type of estimates are very relevant in Harmonic Analysis in order to understand the restriction properties of the Fourier transform to curved surfaces. From that point of view it is

very natural to assume initial conditions $u_0$ such that its Fourier transform $\hat{u}_0$ is in $L^p$. A lot of progress has been done when $\mathcal{L}_0 = \Delta$ but as far as I know there are no results for general $\mathcal{L}_0$ for $p > 2$; see [33], [34], and [25].

Finally let us recall the pseudo-differential operators mentioned in Section 5. As I already said the calculus we develop in [22] and [23] is quite rudimentary, and I think there are many interesting properties to be understood. For example, we do not know if the inequality proved in Proposition 5.1 can be extended to $L^p$ for $p \neq 2$. Another limitation of our approach is that we can construct the integrating factor only for hamiltonians which are free outside of a compact set. It should be enough to assume only that the hamiltonian satisfies hypothesis H2.

## References


[1] Chang, N. H., Shatah, J., Uhlenbeck, K., Schrödinger maps. *Comm. Pure Appl. Math.* **53** (2000), 590–602.

[2] Chihara, H., Local existence for semilinear Schrödinger equations. *Math. Japon.* **42** (1995), 35–51.

[3] Constantin, P., and Saut, J. C., Local smoothing properties of dispersive equations. *J. Amer. Math. Soc.* **1** (1989), 413–446.

[4] Craig, W., Kappeler, T., and Strauss, W., Microlocal dispersive smoothing for the Schrödinger equation. *Comm. Pure Appl. Math.* **48** (1995), 769–860.

[5] Da Rios, L. S. On the motion of an unbounded fluid with a vortex filament of any shape. *Rend. Circ. Mat. Palermo* **22** (1906), 117–135 (in Italian).

[6] Doi, S., On the Cauchy problem for Schrödinger type equations and the regularity of solutions. *J. Math. Kyoto Univ.* **34** (1994), 319–328.

[7] Doi, S., Remarks on the Cauchy problem for Schrödinger–type equations. *Comm. Partial Differential Equations* **21** (1996), 163–178.

[8] Doi, S. , Smoothing effects for Schrödinger evolution equation and global behavior of geodesic flow. *Math. Ann.* **318** (2000), 355–389.

[9] Hasimoto, H. A soliton on a vortex filament. *J. Fluid Mech.* **51** (1972), 477–485.

[10] Hayashi, N., and Ozawa, T., Remarks on nonlinear Schrödinger equations in one space dimension. *Differential Integral Equations* **7** (1994), 453–461.

[11] Ichinose, W., On $L^2$ well-posedness of the Cauchy problem for Schrödinger type equations on a Riemannian manifold and Maslov theory. *Duke Math. J.* **56** (1988), 549–588.

[12] Kato, T., On the Cauchy problem for the (generalized) Korteweg-de Vries equation. In *Studies in applied mathematics* (ed. by V. Guillemin), Adv. Math. Suppl. Stud. 8, Academic Press, New York 1983, 93–128.

[13] Kenig, C. E., Ponce, G., and Vega, L., Oscillatory integrals and regularity of dispersive equations. *Indiana Univ. Math. J.* **40** (1991), 33–69.

[14] Kenig, C. E., Ponce, G., and Vega, L., Small solutions to nonlinear Schrödinger equations. *Ann. Inst. Henri Poincaré* **10** (1993), 255–288.

[15] Kenig, C. E., Ponce, G., and Vega, L., On the Zakharov and Zakharov-Schulman Systems. *J. Funct. Anal.* **127** (1995), 202–234.

[16] Kenig, C. E., Ponce, G., and Vega, L., Smoothing effects and local existence theory for the generalized nonlinear Schrödinger equations. *Invent. Math.* **134** (1998), 489–545.

[17] Kenig, C. E., Ponce, G, and Vega, L., The Cauchy problem for quasi-linear Schrödinger equations. *Invent. Math.* **158** (2004), 343–388.

[18] Kenig, C. E., Ponce, G, and Vega, L., On the ill-posedness of some canonical dispersive equations. *Duke Math. J* **106** (2001), 617–633.

[19] Kenig, C. E., Ponce, G, and Vega, L., On the Cauchy problem for linear Schrödinger systems with variable coefficient lower order terms. In *Harmonic analysis and number theory* (ed. by S. W. Drury and M. Ram Murty), CMS Conf. Proc. 21, Amer. Math. Soc., Providence, RI, 1997, 205–227.

[20] Kenig, C. E., Ponce, G, and Vega, L., On the smoothing properties of some dispersive hyperbolic systems. In *Nonlinear Waves* (Sapporo, 1995), GAKUTO Internat. Ser. Math. Sci. Appl. 10, Gakkōtosho, Tokyo 1997, 221–229.

[21] Kenig, C. E., Ponce, G, and Vega, L., The Initial Value Problem for the General Quasi-linear Schrödinger Equation. Preprint.

[22] Kenig, C. E., Ponce, G, and Vega, L., Variable coefficient Schrödinger flows for ultrahyperbolic operators. *Adv. Math.* **196** (2005), 373–486.

[23] Kenig, C. E., Ponce, G, and Vega, L., The general quasilinear ultrahyperbolic Schrödinger equation. *Adv. Math.*, to appear.

[24] Kruzhkov, D. J., and Faminskii, A. V., Generalized solutions for the Cauchy problem for the Korteweg-de Vries equation. *Math. USSR-Sb.* **48** (1990), 391–421.

[25] Lee, S., Bilinear restriction estimates for surfaces with curvatures of different signs. *Trans. Amer. Math. Soc.* **358** (2006) 3511–3533.

[26] Lim, W-K., and Ponce, G. On the initial value problem for the one dimensional quasilinear Schrödinger equation. *SIAM J. Math. Anal.* **34** (2003), 435–459.

[27] Mizohata, S., *On the Cauchy Problem*. Notes Rep. Math. Sci. Engrg. 3, Academic Press, Orlando, FL, 1985.

[28] Poppenberg, M. Smooth solutions for a class of fully nonlinear Schrödinger type equations. *Nonlinear Anal.* **45** (2001), 723–741.

[29] Rolvung, C., Non-isotropic Schrödinger equations. PhD. Dissertation, University of Chicago, 1998.

[30] Rogers, K., Vargas, A., Vega L., Pointwise convergence of solutions to the nonelliptic Schrödinger equation. *Indiana Univ. Math. J.*, to appear.

[31] Sjölin, P., Regularity of solutions to the Schrödinger equations. *Duke Math. J.* **55** (1987), 699–715.

[32] Sulem, C., Sulem P.L., *The Nonlinear Schrödinger Equation, Self-focusing and Wave Collapse*. Appl. Math. Sci. 139, Springer-Verlag, New York 1999.

[33] Tao, T., A Sharp bilinear restriction estimate for paraboloids. *Geom. Funct. Anal.* **13** (2003), 1359–1384.

[34] Vargas, A., Restriction theorems for a surface with negative curvature. *Math. Z.* 249 (2005), 97–111.

[35] Vega, L., El multiplicador de Schrödinger: la función maximal y los operadores de restricción. PhD Thesis, Universidad Autónoma de Madrid, 1988.

[36] Vega, L., The Schrödinger equation: pointwise convergence to the initial data. *Proc. Amer. Math. Soc.* **102** (1988), 874–878.

[37] Vega, L., Small perturbations of the free Schrödinger equation. In *Recent Advances in Partial Differential Equations* (ed. by M. A. Herrero, E. Zuazua), RAM Res. Appl. Math. 30, John Wiley & Sons, Chichester 1994, 115–130.

Departamento de Matemáticas, Universidad del País Vasco, Apdo. 644, 48080 Bilbao, Spain
E-mail: luis.vega@ehu.es

# Singular solutions of partial differential equations modelling chemotactic aggregation

Juan J. L. Velázquez

**Abstract.** This paper reviews several mathematical results for partial differential equations modelling chemotaxis. In particular, questions like singularity formation for the Keller–Segel model and continuation of the solutions beyond the blow-up time will be discussed. Some of the open problems that remain for the Keller–Segel model as well as some new mathematical problems arising in the study of chemotaxis problems will be discussed.

**Mathematics Subject Classification (2000).** 35K55, 35B40, 92B05.

**Keywords.** Keller–Segel model, singularity formation, chemotaxis, continuation beyond the blow-up.

## 1. Introduction

There are several relevant biological phenomena that involve the type of cell interaction that is known as chemotaxis. This word denotes the capability of many cells to react to chemical stimuli and move towards an increasing or decreasing chemical gradient. Chemotaxis plays a relevant role in biological processes like embryogenesis, angiogenesis or others.

A particular biological process that has deserved considerable attention by biologists, mathematicians and physicists is the phenomenon of chemotactic aggregation. Several unicellular organisms, like Dictyostelium discoideum and Myxococcus xanthus under conditions of environmental stress begin a complex cascade of chemical processes having as a major consequence the release away from the cell of a chemical substance that has chemoattractant properties in the cells themselves. As a consequence, cells begin to approach to each other. This yields to the formation of dense cellular aggregates where the cells usually begin a differentiation process and as a final result the formation of a fruiting body containing cell sporae that remain in such dormant state until they find suitable environmental conditions where they can proliferate again. From the biological point of view an appealing feature of these social organisms is that they expend part of their life cycle as unicellular organisms and the other part as multicellular organisms.

The details of the phenomenon of chemotactic aggregation change very much from organism to organism. On the other hand, even during the simplest stages of the

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

process it is possible to observe many interesting patterns like spiral waves, cell stream formation and others. Nevertheless, in spite of the complexity of this biological process, some of the main features of the problem are simple enough to motivate several mathematicians to derive models that could amount at least for some of the most important features of the phenomenon.

## 2. The Keller–Segel model

The earliest attempt to describe chemotactic aggregation using a system of partial differential equations was the Keller–Segel model that was introduced in [33]. The authors of this model introduced a continuum description of the aggregation process for Dictyostelium discoideum (from now on Dd) containing some of the biological knowledge that had been gained from the experiments made in previous decades. The book [7] contains a great part of the information available on this biological problem at the time of the formulation of the Keller–Segel model.

In a typical aggregation experiments made with Dd many individual amoebae are distributed in the basement of a Petri dish and covered by a liquid layer. Under suitable conditions the cells begin emitting chemical pulses that trigger the aggregation process.

The Keller–Segel model describes this process assuming that there are only two relevant variables in the problem, namely the cell concentration $n$ and the chemical concentration of the substance that propagates the signals between the cells that will be denoted as $c$. Both concentrations are understood to be measured for unit of surface in the basis of the Petri dish. The chemical substance propagating the chemical signals in the case of Dd was identified in the late 60s and it turns out to be the chemical known as cAMP. The functions are assumed to be during the aggregation process functions of the position in the Petri dish $x$ as well as the time $t$, i.e.,

$$n = n(x, t),$$
$$c = c(x, t).$$

The validity of this description requires to measure the functions $n$ and $c$ in a length scale larger than the typical distance between cells that is of the order of some hundreds of microns. Under this assumption it is natural to write the following continuity equations for the densities $n$ and $c$:

$$\frac{\partial n}{\partial t} + \nabla \cdot (j_n) = 0, \tag{1}$$
$$\frac{\partial c}{\partial t} + \nabla \cdot (j_c) = f(n, c), \tag{2}$$

where $j_n$, $j_c$ are the cell fluxes and chemical fluxes respectively. The function $f(n, c)$ describes the production of chemical by the cells as well as the decay of the concentration of $c$ due to its interaction with the substances placed in the extracellular matrix.

It is implicitly assumed in (1) that processes like mitosis or cell death do not play any relevant role. This assumption is reasonable because such processes take place in a time scale much longer than the one related to chemotactic aggregation.

In order to complete model it remains to prescribe the functions $j_n$, $j_c$, $f(n, c)$. Concerning the chemical fluxes the most natural assumption is to assume that the chemical diffuses according to the classical Fick's law:

$$j_c = -D_c \nabla c \tag{3}$$

where $D_c$ is the diffusion coefficient for the chemical.

On the other hand, the Keller–Segel model assumes that the cell motions are the superposition of two effects, namely a random motility and a drift towards the regions having a larger concentration of chemical due to the effect of the chemotaxis. In the case of amoeba-like cells like Dd it is experimentally observed that the cells, that move by means of the expansion and retraction of pseudopods, have some kind of random component in their motion resembling, in a suitable length scale the motion of a brownian particle. On the other hand, it is experimentally observed that the drifting motion of the cells is, on average, proportional to the gradient of chemical concentration. These features make reasonable to assume that the random motility follows the standard Fick's law for diffusive processes and that the drifting motion yields an additional cell flux proportional to $n\nabla c$. The closure relation for the cell flux then becomes

$$j_n = -D_n \nabla n + \chi n \nabla c, \tag{4}$$

where $D_n$ is the diffusion coefficient for the cells and $\chi$ will be termed as chemotactic sensitivity.

Finally, in order to determine the function $f(n, c)$ there are two features that it is important to take into account. The molecules of cAMP degrade with a characteristic life-time, due to their interaction with the molecules of the extracellular membrane. On the other hand, the production of chemical for unit of area is proportional to the cell concentration $n$ if the chemical production of each cell it is assumed to be approximately independent from the others. Under these assumptions the formula for $f(n, c)$ would be

$$f(n, c) = \alpha n - \beta c \tag{5}$$

with $\alpha > 0$, $\beta > 0$.

Combining the equations (1)–(5) the following system of equations follows:

$$\frac{\partial n}{\partial t} = D_n \Delta n - \chi \nabla \cdot (n \nabla c), \tag{6}$$

$$\frac{\partial c}{\partial t} = D_c \Delta c + \alpha n - \beta c. \tag{7}$$

The system (6), (7) is a particular case of classical Keller–Segel model that was introduced in [33]. Using a suitable set of dimensionless variables, it is possible to

reduce (6), (7) to the analysis of the particular case:

$$\frac{\partial n}{\partial t} = \Delta n - \chi \nabla \cdot (n \nabla c), \tag{8}$$

$$\frac{\partial c}{\partial t} = L \Delta c + n - \beta c. \tag{9}$$

This system is usually solved in a domain $\Omega \subset I\!R^2$ for positive times $t > 0$ with suitable initial data $n(x, 0) = n_0(x) \geq 0, c(x, 0) = c_0(x) \geq 0$ and zero flux boundary conditions:

$$\frac{\partial n}{\partial \nu} - \chi n \frac{\partial c}{\partial \nu} = 0, \quad \frac{\partial c}{\partial \nu} = 0, \quad x \in \partial\Omega, \quad t > 0 \tag{10}$$

where $\nu$ is the outer normal at the boundary $\partial\Omega$. Notice that, under these boundary conditions, the total number of cells $\int_\Omega n_0(x)\, dx \equiv N_0$ is conserved during the evolution of the system. In several of the discussions below it will be assumed that $N_0$ is a real number of order one, something that at a first glance might look strange for a number that denotes the rather large number of cells contained in $\Omega$. However, this is due only to the fact that in the formulation (8), (9) dimensionless variables have been used. A number $N_0$ of order one for the solutions of the system (8), (9) is in reality a huge number of cells if the dimensional form of the equations (6), (7) is used.

The problem (8), (9) turned out to be a source of interesting mathematical problems. From the mathematical point of view the most interesting feature of the system (8), (9) is the nonlinear term $\chi \nabla \cdot (n \nabla c)$. If the chemotactic interaction between cells is chemoattractive, i.e. if $\chi > 0$, this term yields singularity formation in finite time. The peculiar form of this nonlinear term is a rather common feature of the chemotaxis models. The study of the consequences of this term in the dynamics of the solutions of (8), (9) has led to the development of several mathematical tools by different authors.

## 3. Singularity formation in chemotaxis models

Childress suggested that the solutions of (6), (7) could generate singularities and that the process of chemotactic aggregation could be thought as the formation of a singularity (cf. [13]). The first rigorous proof of blow-up in a chemotaxis model was obtained by Jäger and Luckhaus in [31]. These authors took advantage of the fact that the diffusion coefficient for the chemical $D_c$ is much larger than the diffusion coefficient $D_n$ for the cells. In that particular limit the system of equations (8), (9) with nonzero flux boundary conditions can be reduced to the simpler problem

$$\frac{\partial n}{\partial t} = \Delta n - \chi \nabla \cdot (n \nabla c), \tag{11}$$

$$0 = L \Delta c + n - \bar{n}, \tag{12}$$

where $\bar{n} = \frac{1}{|\Omega|} \int_\Omega n\, dx$. Notice that in order to solve (11), (12) only the initial data $n(x, 0) = n_0(x)$ must be prescribed.

Jäger and Luckhaus obtained two basic results for the solutions of (11), (12) that established the basic framework for many of the subsequent researches in this type of problems. They proved that the solutions of (11), (12) with $N_0$ small are globally bounded. On the other hand, [31] contains also a large class of radial initial data $n_0(x)$ for which the corresponding solution of (11), (12) becomes unbounded in finite time, or using the standard terminology used by the partial differential community, the solutions of (11), (12) blow up in finite time for suitable initial data.

Notice that the global existence result in [31] implies that there exists a threshold for the number of cells $N_0$, below which the solutions of (11), (12) do not exhibit singularities in finite time. There have been several researches trying to compute the value of such threshold number. In the case of radial solutions, it was proved by Nagai (cf. [37]) that the smallest number of cells needed to have blow up in finite time is $8\pi/\chi$. In nonradial cases the threshold for the mass is $4\pi/\chi$ (cf. [6], [16]).

## 4. Chemotactic aggregation

The blow up results mentioned in the previous section do not imply that the solutions of the system (11), (12) develop a Dirac mass in a finite time. The onset of such Dirac mass at the time of formation of the singularity $t = T$ seems the most natural outcome, as the previous discussion concerning the existence of a threshold for the number of cells needed to create a singularity suggests. However, the derivation of such conclusion is not so obvious. Indeed, a possibility that cannot be excluded in principle is the formation of a singularity where the number of cells contained in a small ball near the point where the singularity appears would oscillate infinitely often without converging to any number. Another possibility that is not easy to rule out in nonradial cases is the existence of a family of balls containing a large fraction of the total number of cells whose diameter decreases to zero as $t \to T^-$ and whose centers move erratically by the domain $\Omega$.

We will denote as chemotactic aggregation the formation of a Dirac mass at a finite time $t = T < \infty$. In [17] was obtained a class of solutions yielding chemotactic aggregation. Moreover, for such a solutions there was a detailed description of the asymptotic behaviour of the solutions near the singularity. These solutions satisfy

$$n(\cdot, t) \rightharpoonup \frac{8\pi}{\chi}\delta(\cdot) + f(\cdot) \quad \text{as } t \to T^-, \tag{13}$$

where

$$f(x) \sim \frac{8e^{-(\gamma+2)}}{|x|^2} e^{-2\sqrt{|\log(|x|)|}}(1 + o(1)) \quad \text{as } |x| \to 0, \tag{14}$$

$\gamma$ being the classical Euler constant. Moreover:

$$n(x,t) \sim \frac{8}{(T-t)(\varepsilon(|\log(T-t)|))^2} \frac{1}{\left(1 + \frac{|x|^2}{(T-t)(\varepsilon(|\log(T-t)|))^2}\right)^2} \quad \text{as } t \to T^- \tag{15}$$

for $|x| \leq C\sqrt{T-t}\varepsilon(|\log(T-t)|)$, where

$$\varepsilon(\tau) \sim 2e^{-\frac{2+\gamma}{2}} e^{-\sqrt{\frac{\tau}{2}}} \left(1 + O\left(\frac{\log(\tau)}{\sqrt{\tau}}\right)\right) \quad \text{as } \tau \to \infty. \tag{16}$$

It is interesting to point out that these solutions are not self-similar solutions in the sense that such term is usually understood. The most common meaning that it is given to the term self-similar solutions is the one of solutions that are invariant by a group of symmetries, and most often by a group of rescalings. If the term $\overline{n}$ that gives a low order contribution is ignored in the equations (11), (12) the resulting equations are invariant under the rescaling group:

$$x \to \lambda x, \quad t \to \lambda^2 t, \quad n \to \frac{1}{\lambda^2} n, \quad c \to c, \tag{17}$$

where $\lambda$ is an arbitrary positive number. It can be shown that at least in the radial two dimensional case there are not self-similar solutions of (11), (12), even if the term $\overline{n}$ is neglected (cf. [20]). On the other hand, it is not hard to see that solutions with the asymptotics (15) are not invariant under the rescaling group (17) due to the presence of the terms $e^{-\sqrt{\frac{|\log(T-t)|}{2}}}$ in (16) that are not power laws. More precisely

$$(T-t)^a \ll e^{-\sqrt{\frac{|\log(T-t)|}{2}}} \ll 1 \quad \text{as } t \to T^-$$

for any $a > 0$. In the terminology of applied mathematicians such terms are often called "logarithmic corrections", even if they are not strictly logarithmic functions.

The computation of this logarithmic corrective term is the main technical difficulty solved in [17]. The key idea used in that paper was to derive first an approximation of the solution near the blow-up time using the so-called "matched asymptotic expansions". Such expansions are a heuristic, non fully mathematically rigorous procedure of deriving approximated formulae for the solutions different types of equations that contain large or small parameters. These methods are widely used in many fields applied mathematics, often combined with numerical simulations that provide an independent test of their validity. The basic idea of this method consists in to compute perturbative series for the solutions of the equations under consideration by simpler equations, something that is possible due to the presence of large or small parameters in the problem. However, such approximations of the solutions often lose their validity in some regions of the space because the form of the obtained solutions determines that some of the terms that had been previously ignored become important in some particular areas of the space of parameters. The solutions are then analysed in these specific regions introducing suitable rescalings and changes of variables that are often suggested by the form of the approximate solutions previously computed. The resulting equations can then be also analysed in a perturbative manner and in this way the form of the solutions in such new variables can be obtained too. In order to assert the validity of the obtained formula it remains to check that both obtained approximated

solutions agree in a region of common validity. This agreement is usually termed as "matching".

In the study of (11), (12) the small parameter is the distance between the time variable and the blow-up time, i.e. $(T-t)$. The analysis of the solutions was made decomposing the space of independent variables $(x,t)$ in three different regions, namely:

$$\begin{aligned} &\text{(A)} \quad |x| \ll \sqrt{T-t}, \\ &\text{(B)} \quad |x| \approx \sqrt{T-t}, \\ &\text{(C)} \quad |x| \gg \sqrt{T-t}. \end{aligned}$$

The onset of the parabolic rescaling $\sqrt{T-t}$ it is very natural due to the parabolic character of the system (11), (12). In each of these regions these equations can be approximated to the leading order by a different type of equation whose solution can be obtained in an explicit manner. More precisely, the derived solutions solve a nonlinear ordinary differential equation in the region (A), a linear parabolic equation that reduces to the heat equation in the region (B), and a quasilinear hyperbolic equation in the region (C). The corrective term $e^{-\sqrt{\frac{|\log(T-t)|}{2}}}$ was obtained setting that the width of the region where the aggregating mass is concentrated is an unknown function $\varepsilon(t)$ whose precise form is computed matching the solutions obtained in the regions (A) and (B). Such matching condition provides an integro-differential equation for $\varepsilon(t)$ that allows to compute the "logarithmic correction" $e^{-\sqrt{\frac{|\log(T-t)|}{2}}}$. The details of this formal computation can be found in [18] and also in [45], [46].

The rigorous construction of the solutions whose formal description is given above was made reducing the problem to the one of finding the zeroes of a finite dimensional problem. This was achieved choosing an initial class of initial data depending on a finite number of parameters and showing that the choice of such parameters that solves a suitable equation provide some initial data whose corresponding solution blows up at the time $t = T$ with the asymptotic behaviour computed in an heuristic manner before. The details of the argument can be found in [17]. A crucial point in the argument is to use the formal asymptotics of the solutions as a guide to derive suitable "a priori" estimates for the solutions of (11), (12). This argument have been used in the construction of solutions with a prescribed blow-up behaviour in many other problems (cf. for instance [3], [10]).

All the previous analysis was made for the simplified version of the Keller–Segel model introduced by Jäger and Luckhaus. Nevertheless, the same results can be obtained for the whole Keller–Segel system (8), (9) (cf. [22]). Some technical difficulties arise due to the fact that for radial solutions it is possible to reduce (11), (12) to the study of a scalar equation. Such reduction is not possible in the case of the whole system (8), (9).

It is worth mentioning that this study provides a simple formula, originally derived in [13], relating the different parameters from the Keller–Segel model and the number

of aggregating cells. Indeed, rewriting (13) with the original dimensional variables it follows that the number of cells aggregating for the solutions of (8), (9) is

$$N_{\text{aggr.cells}} = \frac{8\pi D_c D_n}{\alpha \chi}.$$

## 5. Some analogies between chemotactic aggregation and the melting of ice balls

It is interesting to remark that the same type of "logarithmic correction" that has been described above (cf. (15)) appears in other problem that at a first glance looks rather different from the problem of chemotactic aggregation. Suppose that one tries to describe the size of a melting ice cylinder immersed in a big reservoir of water. One usual way of describing such process is by means of the so-called Stefan problem. In the resulting model it is assumed that the heat transfer in both the ice and the water follows the classical Fourier's law. Therefore, in both phases the temperature satisfies a heat equation. On the other hand, in the interface separating both phases the temperature takes a constant value that is the melting temperature of the water at the value of the pressure that the experiment is made. This assumption is not completely true if surface tension effects are taken into account, but these effects are relevant only for very small radii of the cylinder and therefore they can be ignored during most of the process. Moreover, we will assume also that the heat conductivity of the ice is much higher than the one of the ice, since this makes the problem easier to analyse and it does not change the final conclusions. A final feature that must be incorporated in the model is the fact that the melting of a given volume of ice requires to provide to it the amount of energy known as latent heat. The mathematical formulation of this condition provides an equation for the motion of the interface separating the ice and the liquid water. The resulting model, in dimensionless units and in the radial case is the following:

$$\frac{\partial \theta}{\partial t} = \Delta \theta \quad x \in I R^2, \quad |x| > R(t), \quad t > 0, \tag{18}$$

$$\theta = \theta_m, \quad |x| = R(t), \quad t > 0, \tag{19}$$

$$\dot{R}(t) = -\frac{\partial \theta}{\partial r}(R(t), t), \tag{20}$$

where $\theta(x, t)$ is the temperature of the liquid water. In absence of undercooled water we must assume that $\theta(x, 0) = \theta_0(x) \geq 0$.

This problem has classical solutions for a large class of initial data $\theta_0(x)$. Let us suppose also that $\theta_0(x) \to \theta_\infty > 0$ as $|x| \to \infty$. It might be seen that for such data the radius of the ice ball $R(t)$ decreases and eventually disappears in finite time. At such time a singularity arises for this free boundary problem.

As in the case of the Keller–Segel model there are not self-similar solutions describing this singularity. A description of this singularity using formal asymptotic

expansions was obtained in [39]. It turns out that the radius of such balls near the time $t = T$ for the vanishing of the spheres is given by:

$$R(t) \sim C\sqrt{T-t}e^{-\frac{\sqrt{2}}{2}|\log(T-t)|^{\frac{1}{2}}} \quad \text{as } t \to T$$

Asymptotic expansions for the solutions of the same problem in the non radial case were obtained in [19]. For these solutions the interface behaves asymptotically as an ellipsoidal cylinder near the time of the vanishing of the ice.

## 6. Some results for the Jäger–Luckhaus model in three dimensions

There are biological situations where it makes sense to analyse the three dimensional version of the Keller–Segel. A relevant example is the study of the aggregates of the bacteria E.coli.

From the mathematical point of view the type of singularities arising for the Keller–Segel model are very different in the three dimensional case and in the two dimensional case. It is possible to construct singular solutions blowing up in a line, just adding an additional dimension to the solution behaving as in (13)–(15). On the other hand, in three dimensions there exist radial self-similar solutions that yield singularities in a finite time without mass aggregation. On the other hand in three dimensions there exist a mechanism of chemotactic aggregation that is rather different from the one previously described for the two dimensional case. Such aggregation mechanism is driven by the first order terms in (11), except for a small boundary layer where the diffusive term $\Delta c$ becomes essential. For these solutions the mass is concentrated as $t \to T^-$ in a layer placed at a distance of order $(T-t)^{1/3}$. The detailed description of such solutions can be found in [9], [20], [21].

## 7. On the continuation of the solutions beyond the blow-up Time

In the last decade several models have been suggested in order to describe the effects that could stop the aggregation process in different organisms. Several biochemical processes that could stop the aggregation of E. coli if the cell density reach high values were described in [8]. Another partial differential equations that stop cell aggregation were considered in [25], [46], [47]). It was assumed in [25] that the cell velocity vanishes for high cell concentrations. In [46] was assumed also that the cell velocity decreases also with the velocity for high values of the concentration. More precisely, the model studied in [46], [47] was the following:

$$\frac{\partial n}{\partial t} = \Delta n - \nabla \cdot (g_\varepsilon(n)\nabla c), \tag{21}$$

$$0 = \Delta c + n, \tag{22}$$

where

$$g_\varepsilon(n) = \frac{1}{\varepsilon} Q(\varepsilon n), \quad \varepsilon > 0, \tag{23}$$

and $Q$ is an increasing function satisfying

$$Q(s) \sim s - \alpha s^2 \quad \text{as } s \to 0,$$
$$Q(s) \sim L > 0 \quad \text{as } s \to \infty.$$

The solutions of (23) are globally bounded for each $\varepsilon > 0$. On the other hand, the model (21)–(23) converges formally as $\varepsilon \to 0$ to the model

$$\frac{\partial n}{\partial t} = \Delta n - \nabla \cdot (n \nabla c), \tag{24}$$
$$0 = \Delta c + n, \tag{25}$$

that in two dimensions might yield chemotactic aggregation in finite time. It would be then natural try to understand the asymptotics of the solutions of (21)–(23) for arbitrary times as $\varepsilon \to 0$. Notice that the number of cells $\int_\Omega n\, dx$ remains constant for the solutions of (21)–(23). Therefore, even if the solutions of (21)–(23) become unbounded, the solutions of these equations should not become unbounded everywhere. The study of the dynamics of the solutions of (21)–(23) as $\varepsilon \to 0$ was made in [46], [47] using formal matched asymptotic expansions. The conclusion of such analysis was that it is possible to obtain asymptotic expansions valid in all the regions of the space for some solutions of (21)–(23) that behave asymptotically as $\varepsilon \to 0$ as

$$n(x,t) = \sum_{i=1}^{N} M_i(t) \delta(x - x_i(t)) + n_{\text{reg}}(x,t),$$

where $n_{\text{reg}}(x,t)$ is a bounded function. Moreover, the functions $M_i(t), x_i(t), n_{\text{reg}}(x,t)$ satisfy the following problem:

$$\frac{\partial n_{\text{reg}}}{\partial t} = \Delta n_{\text{reg}} + \sum_{j=1}^{N} \frac{M_j(t)}{2\pi} \frac{(x - x_j(t))}{|x - x_j(t)|^2} \cdot \nabla n_{\text{reg}} - \nabla (n_{\text{reg}} \nabla c_{\text{reg}}), \tag{26}$$

$$c_{\text{reg}} = -\frac{1}{2\pi} \int_{I R^2} \log(|x - y|) n_{\text{reg}}(y,t)\, dy, \tag{27}$$

$$\dot{x}_i(t) = \Gamma(M_i(t)) A_i(t), \quad i = 1, \ldots, N, \tag{28}$$

$$A_i(t) = -\sum_{j=1}^{N} \frac{M_j(t)}{2\pi} \frac{(x - x_j(t))}{|x - x_j(t)|^2} + \nabla c_{\text{reg}}(x_i(t), t), \tag{29}$$

$$\frac{dM_i(t)}{dt} = c_{\text{reg}}(x_i(t), t) M_i(t), \quad i = 1, \ldots, N, \tag{30}$$

where $\Gamma(\cdot)$ is a positive function defined for values of its argument larger than $8\pi$.

The problem (26)–(30) can be considered as a moving boundary problem. These equations indicate that there exist solutions of (21)–(23) having some concentration regions where the cells accumulate and that interact between themselves and with the cells away from the aggregates.

The solvability of the problem (26)–(30) is not entirely obvious due to the motion of the points $x_i(t)$ as well as the presence of the terms $\frac{M_j(t)}{2\pi}\frac{(x-x_j(t))}{|x-x_j(t)|^2}\cdot\nabla n_{\text{reg}}$ in (26). The local well-posedness of (26)–(30) in Hölder spaces has been proved in [48].

In [47] has been obtained using matched asymptotic expansions a description of the way in which the saturation of the chemotactic attraction for high values of the concentration stops the aggregation process and yields the formation of a concentration region for the density.

## 8. Some open questions for the Keller–Segel model

In the last two decades there have been several relevant advances in the understanding of the Keller–Segel model. There are, however, still many unsolved questions that could pose challenging analysis problems. I will describe shortly some of the ones that in my opinion are more relevant.

Probably, the most important problem that remains in order to understand completely the blow-up for the Keller–Segel model is to show that for arbitrary two dimensional domains $\Omega$ and arbitrary initial data, all the blowing up solutions of (8), (9), (or the simplified version (11), (12)) converge locally near the blow-up to a Dirac mass.

A more ambitious version of this problem would be to show that all the solutions that blow up in finite time behave near the singularity as indicated in (13)–(15). Experts in blow-up would immediately argue here that, since the solutions with large amount of cells blow up and the solutions with a small number of cells do not blow-up, there exists a transition regime between the one associated to global existence and the one associated to blow-up in finite time. I think that the most spread opinion among the mathematicians working in the Keller–Segel model about this point is that the transition regime corresponds precisely to the solutions having a total number of cells of $8\pi/\chi$, and that this critical amount of cells should lead to the type of behaviour that it is usually known as "blow-up in infinity time" (cf. for instance [29]). This would mean that the solutions would be globally defined but the solutions of (8), (9), (or (11), (12)) should eventually approach to a Dirac mass as $t \to \infty$. The blow-up mechanism (13)–(15) might be obtained with any number of cells strictly greater $8\pi/\chi$ cells. For the critical number of cells, the description of the long time asymptotics as $t \to \infty$ has not been obtained even at the formal level.

A problem that could shed some light in the question of finding a complete classification of the singular behaviours for the Keller–Segel model is the study of the stability of the solutions of (11), (12) with the behaviour (13)–(15). In principle this

problem looks more amenable to analysis because it reduces to the study of a local problem. This stability study for these particular solutions has been made in [45] using formal computations linearizing formally around the solution obtained in [17]. This kind of linearization is customarily made by applied mathematicians working in problems that involve blow-up phenomena. Nevertheless, the study of such stability is more involved that in many of the blow-up problems so far considered due to the involved structure of boundary layers that is needed to describe the solution behaving as in (13)–(15). To prove in a fully rigorous manner that the solutions of (11), (12) with the behaviour (13)–(15) would require, most likely, to make fully rigorous the arguments in [45], something that would require to study in detail several parabolic problems described in [45]. A similar analysis challenge is the one posed by the proof in a fully rigorous manner of the results concerning "continuation beyond blow-up" mentioned in Section 7.

There is a huge wealth of problems associated to the study of the singularities for the Keller–Segel or the Jäger–Luckhaus problems in three spatial dimensions. As indicated in Section 6 in this case there are many more singular behaviours, whence a complete classification of blowing up solutions seems much harder. It is interesting to point out that the solutions of the Jäger–Luckhaus model blowing up in a line that are obtained adding an additional dimension to the solutions blowing up as in (13)–(15) seem to be unstable under nonconstant perturbations along that line, as the formal computations in [5] suggest.

Let us finally remark that there seem to be several analogies between the Keller–Segel model (or the Jäger–Luckhaus approximation) and the classical Stefan problem. This is particularly clear in the results mentioned in Section 5, but there are other points in the mathematical analysis of both problems where these analogies can be seen. It is not unlikely that the Stefan problem could be derived, at least formally, as a suitable asymptotic limit of the Keller–Segel model, in the same form as the Stefan problem and many other related free boundary problems can be derived from the so-called phase field limits. If the connection between these problems is found, it would be perhaps possible to explain the analogies found in Section 5. It would probably be possible also to use the large amount of information available for the Stefan problem in order to describe the behaviour of some class of initial data for the Keller–Segel model.

In this paper the description and stability of the steady states solutions of the Keller–Segel model has not been considered. There are several results concerning the structure of the steady states of this problem (cf. [40]). A detailed review of the different mathematical results available in the literature for the stationary and the evolutionary Keller–Segel model can be found in [27], [28].

## 9. Beyond the Keller–Segel model

All the previous discussion has focused exclusively in the study of the Keller–Segel model. As it was explained in Section 2 the Keller–Segel model is a continuous approximation of a rather complicated aggregation process. In recent decades the study of the process of chemotactic aggregation has developed in many more directions than in the study of this specific model. In this section, I will describe briefly some of this researches to illustrate the type of mathematical problems that have arisen in the study of this biological process.

One of the research directions that has deserved great attention and that was originated by the papers [1], [15], [38] is the study of kinetic or stochastic models describing the cell dynamics.

The idea underlying the stochastic models is to describe the dynamics of each individual cell using a stochastic differential equation. The information contained in the differential equation is that cells move in a rectilinear manner at constant speed during some time intervals. At the end of such intervals the direction of the velocity changes in a random manner. In order to obtain a chemotactic dynamics the models introduce some bias towards the regions having greater chemical concentration, something that can be made in several different ways. One possibility is to assume that the rate of change in the direction of motion is a function on the change of concentration of chemical. Another different possibility is to assume that the new direction of motion is biased towards the direction of largest chemical concentration. The first possibility is motivated by the well studied motion mechanism of E. coli that takes place by means of different types of discrete jumps in the space known as "runs" and "tumbles" (cf. the description in [4]). The second one can be thought as a reasonable approximation to the dynamics of amoebae-like cells like Dd. In order to avoid introducing in this model direct cell-cell interactions it must be assumed that the mean free path between jumps is much smaller than the cell distance. If it is assumed, in addition, that the characteristic distance for the chemical variation are much larger than the cell distance it might be seen that the particle distributions have small correlations and therefore they might be approximated as the product of one-particle distributions in the space of velocities and positions having the form $f(x, v, t)$. In all the mentioned asymptotic limits it is possible to approximate the evolution equation for the one-particle distribution function by means of the kinetic equation

$$f_t + v\nabla_x f = \int [T(x, v, w) f(x, w, t) - T(x, w, v) f(x, v, t)]\, dw. \tag{31}$$

The transition kernel $T(x, w, v)$ contains the bias towards higher concentrations, and therefore it depends in general in quantities like the chemical concentration, of its time or space derivatives.

In recent years there have been obtained several results proving that in some suitable asymptotic limits the solutions of (31) converge to the solutions of the Keller–Segel system (cf. [12], [23], [24], [30]). Readers familiar with gas-dynamics would

realize that the main assumption in these studies is that the mean free path between jumps is much smaller than the characteristic length associated to the chemical concentration. In all these studies the cell concentration is given by $n(x,t) = \int f(x,v,t)\,dv$. A more direct study of a system of stochastic differential equations that are coupled only through the concentration of the chemical was made in [42]. Nevertheless in this paper was also assumed that the distance between particles is small compared with the characteristic length associated to the chemical, and therefore the correlations between particle distributions are also small.

These results point out to some of the new possible directions for the development of the kinetic (or stochastic) theory of cell motion. Given the huge number of different situations that can arise in the study of cell interactions in biological situations it would be relevant to study the dynamics of stochastic equations in limits where the particle correlations could play a relevant role. The study of such problems, at least in biological problems is largely open.

On the other hand, all the studies of cell dynamics using stochastic or kinetic models described above assume that the cells are separated enough from each other to make cell-cell interaction effects negligible. However in many phenomena of chemotactic aggregation this hypothesis fails at least during some part of the process. For instance, in the case of Dd the cells become at some point a dense aggregate package. Even before reaching that state the cells distribute in some cell streams that cannot be described using a simple model as Keller–Segel, but that had been explained to be due to the instabilities of planar fronts for some more complicated reaction-diffusion systems (cf. [26]). Concerning the aggregate state there have been several attempts to model such cellular state by means of different approaches. I will not try to describe in detail all the results that have been obtained in this extensive research area, but I will mention a few results that could describe some of the main ideas that are been used to study this problem.

An approach that is rather popular in the field of mathematical biology is the use of cellular automata models, often in lattices, having a dynamic that mimics the laws of cell motion. Using this approach it is possible to obtain numerical simulations that very often resemble very much the patterns observed in biological systems. In the specific case of Dd the most remarkable results in this direction are those of [35]. These numerical simulations were able to reproduce the whole life cycle of Dd, including the aggregation, the formation of cell mounds, and the development of the fruiting body. The main difficulty with this approach is that the relation between the parameters in the cellular automata and the biochemical parameters is not an obvious one. On the other hand, the evolution rules that are used in the model are not true mechanical or chemical equations. In any case this approach is providing some insight in several biological problems about the type of interactions that must be taken into account to obtain some specific patterns.

There have been introduced in the literature several models that describe the mechanical interactions of a huge number of cells. One of the most recent papers in this direction, that is probably the one that includes the currently available information

about the mechanical interactions of the cells in a more careful manner is [14]. These results provide some interesting insights on the mechanics of dense cell aggregates. Nevertheless the mathematical object that a specialist in partial differential equations would like to have is a system of continuum equations for the cell aggregates, with a solid basis in physics and chemistry that could play a role in tissues analogous to the one played by the Navier–Stokes equations in fluid mechanics. Actually the Navier–Stokes equations have been used to model the evolution of cell aggregates (cf. [11]) and there has been obtained numerically good qualitative agreements with the experimentally observed patterns (cf. [41]). However, the mechanical properties of dense cell aggregates are most likely rather different from the properties of newtonian fluids.

I will mention shortly another problem related to the aggregation of Dd yielding also partial differential equations problems. This is the process of transmission of chemical signals between cells that yield the aggregation process. In the Keller–Segel model it is assumed that the production of chemical is proportional to the cell density and that the cell velocity is proportional to the concentration of chemical. However, this assumptions are just a simplification of a rather complex process. A more realistic picture of the production of chemical is provided by the theory of excitable systems. According to this picture the cells, upon the arrival of a diffusive chemical wave produce some additional chemical that compensates in this way the effect of its spontaneous degradation. After this production the cell enters in a refractory state lasting a few minutes during which the cell is unable to release chemical. There exists a huge mathematical theory for reaction-diffusion systems whose dynamics has these ingredients (cf. for instance [32]). Systems having such dynamics exhibit a large class of patterns, like travelling waves, spiral waves and several others. Actually, the chemical signalling in Dd aggregates was one of the problems that motivated the development of the theory of dynamical systems. In recent decades there have been several papers introducing models for the cell signalling process that include more detailed information about the biochemical processes taking place in the cell. Some of the most popular models are the ones in [36] and [43]. These models are reaction-diffusion models that are able to reproduce some of the features observed in the experiments that measure the chemical produced by cell aggregates (cf. [44]). Recently, in the article [34] several analytic formulae for magnitudes like the wave velocity, the chemical concentrations and other related quantities were computed using asymptotic methods.

**Acknowledgements.** This paper was partially written during a sabbatical stay at the Max Planck Institute for the Mathematics in the Natural Sciences. The author is partially supported by a fellowship of the Humboldt foundation and by DGES Grant MTM2004-05634. I wish to thank S. Luckhaus for introducing me to the study of the Keller–Segel model. Several of the results described in this paper were obtained in collaboration with M. A. Herrero. I want to also acknowledge several illuminating discussions concerning chemotaxis and mathematical biology with A. Stevens.

## References


[1] Alt, W., Biased random walk models for chemotaxis and related diffusion approximations. *J. Math. Biol.* **9** (1980), 147–177.

[2] Andreucci, D., Herrero, M. A., and Velázquez, J. J. L., The classical one-phase Stefan problem: A catalogue of interface behaviours. *Surveys Math. Indust.* **9** (4) (2001), 247–337.

[3] Angenent, S. B., and Velázquez, J. J. L., Degenerate neckpinches in mean curvature flow. *J. Reine Angew. Math.* **482** (1997), 15–66.

[4] Berg, H., Random walks in biology, Princeton University Press, Princeton, NJ, 1993.

[5] Betterton, D., and Brenner, M. P., Collapsing bacterial cylinders. *Phys. Rev. E* **64** (2001), 519–534.

[6] Biler, P., Local and global solvability of some parabolic systems modeling chemotaxis. *Adv. Math. Sci. Appl.* **8** (1998), 715–743.

[7] Bonner, J. T., *The cellular slime molds*. Princeton University Press, Princeton, NJ, 1967.

[8] Brenner, M. P., Levitov, L. S., and Budrene, E. O., Physical mechanisms for chemotactic pattern formation by bacteria. *Biophys. J.* **74** (1998), 1677–1693.

[9] Brenner, M. P., Constantin, P., Kadanoff, L. P., Schenkel, A., and Venkataramani, S. C., Diffusion, attraction and collapse. *Nonlinearity* **12** (1999), 1071–1098.

[10] Bressan, A., Stable blow-up patterns. *J. Differential Equations* **98** (1992), 57–75.

[11] Bretschneider, T., Vasiev, B., and Weijer, C. J., A model for Dictyostelium slug movement. *J. Theor. Biol.* **199** (1999), 125–136.

[12] Chalub, F. A. C. C., Markowitch, P., Perthame, B., Schmeiser, C., On the derivation of drift-diffusion model for chemotaxis from kinetic equations. ANUM Preprint 14/02, Vienna Technical University 2002.

[13] Childress, S., Chemotactic collapse in two dimensions. In *Modelling of Patterns in Space and Time*, Lecture Notes in Biomath. 55, Springer-Verlag, Berlin 1984, 61–68.

[14] Dallon, J. C., and Othmer, H. G., How cellular movement determines the collective force generated by the Dyctiostelium discoideum slug. *J. Theor. Biol.* **231** (2004), 203–222.

[15] Dickinson, R. B., and Tranquillo, R. T., A Stochastic Model for Cell Random Motility and Haptotaxis Based on Adhesion Receptor Binding Fluctuations. *J. Math. Biol.* **31** (6) 563-600 (1993).

[16] Gajewski, H., and Zacharias, K., Global behaviour of a reaction-diffusion system modelling chemotaxis. *Math. Nachr.* **195** (1998), 77–114.

[17] Herrero, M. A., and Velázquez, J. J. L., Singularity patterns in a chemotaxis model. *Math. Ann.* **306** (1996), 583–623.

[18] Herrero, M. A., and Velázquez, J. J. L., Chemotactic collapse for the Keller–Segel model. *J. Math. Biol.* **35** (1996), 177–194.

[19] Herrero, M. A., and Velázquez, J. J. L., On the melting of ice balls. *SIAM J. Math. Anal.* **1** (1997), 1–32.

[20] Herrero, M. A., Medina, E., and Velázquez, J. J. L., Self-similar blow-up for a reaction-diffusion system. *J. Comp. Appl. Math.* **97** (1–2) (1998), 99–119.

[21] Herrero, M. A., Medina, E., and Velázquez, J. J. L., Finite time aggregation into a single point in a reaction-diffusion system. *Nonlinearity* **10** (1997), 1739–1754.

[22] Herrero, M. A., and Velázquez, J. J. L., A blow-up mechanism for a chemotaxis model. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **24** (1997), 633–683.

[23] Hillen, T., and Othmer, H. G., The diffusion limit of transport equations derived from velocity jump processes. *SIAM J. Appl. Math.* **61** (2000), 751–775.

[24] Hillen, T., and Othmer, H. G., Chemotaxis equations from the diffusion limit of transport equations. *SIAM J. Appl. Math.* **62** (2002), 1222–1250.

[25] Hillen, T., and Painter, K., Global existence for a parabolic chemotaxis model with prevention of overcrowding. *Adv. in Appl. Math.* **26** (2001), 280–301.

[26] Höfer, T., Sherrat, J. A., and Maini, P. K., Cellular pattern formation during Dictyostelium aggregation. *Physica D* **85** (1995), 425–444.

[27] Horstmann, D., From 1970 until present: The Keller–Segel model in chemotaxis and its consequences I. *Jahresber. Deutsch. Math.-Verein.* **105** (3) (2003), 103–165.

[28] Horstmann, D., From 1970 until present: The Keller–Segel model in chemotaxis and its consequences II. *Jahresber. Deutsch. Math.-Verein.* **106** (2) (2004), 51–69.

[29] Horstmann, D., and Wang, G., Blow-up in a chemotaxis model without symmetry assumptions. *Eur. J. Appl. Math.* **12** (2001), 159–177.

[30] Hwang, H. J., Kang, K., and Stevens, A., Global solutions of nonlinear transport equations for chemosensitive movement. *SIAM J. of Math. Analysis* **36** (4) (2005), 1177–1199.

[31] Jäger W., and Luckhaus, S., On explosions of solutions to a system of partial differential equations modelling chemotaxis. *Trans. Amer. Math. Soc.* **329** (1992), 819–824.

[32] Keener, J. P., A geometric theory for spiral waves in excitable media. *SIAM J. Appl. Math.* **46** (1986), 1039–1056.

[33] Keller, E. F., and Segel, L. A., Initiation of slime mold aggregation viewed as an instability. *J. Theor. Biol.* **26** (1970), 399–415.

[34] Litcanu, G., and Velázquez, J. J. L., Singular perturbation analysis of cAMP signalling in Dictyostelium discoideum aggregates. *J. Math. Biol.*, to appear.

[35] Maree, A. F., and Hogeweg, P., Modelling dictyostelium discoideum morphogenesis: the culmination. *Bull. Math. Biol.* **64** (2) (2002), 327–353.

[36] Martiel, J. L., and Goldbeter, A., A model based on receptor desensitization for cyclic AMP signalling in Dictyostelium cells. *Biophys. J.* **52** (1987), 807–828.

[37] Nagai, T., Blow-up of radially symmetric solutions to a chemotaxis system. *Adv. Math. Sci. Appl.* **5** (1995), 581–601.

[38] Patlak, C. S., Random walk with persistence and external bias. *Bull. Math. Biol. Biophysics* **15** (1953), 311–338, .

[39] Riley, D. S., Smith, F. T., and Poots, G., The inward solidication of spheres and circular cylinders. *Int. J. Heat and Mass Transfer* **17** (1974), 1507–1516.

[40] Schaaf, R., Stationary solutions of chemotaxis systems. *Trans. Amer. Math. Soc.* **292** (1985), 531–556.

[41] Siegert, F., and Weijer, C. J., Three-dimensional scroll waves organize Dictyostelium slugs. *Proc. Natl. Acad. Sci. USA* **89** (1992), 6433–6437.

[42] Stevens, A., The derivation of chemotaxis equations as limit of moderately interacting stochastic many particle systems. *SIAM J. Appl. Math.* **61** (2000), 183–212.

[43] Tang, Y., and Othmer, H. G., Excitation, oscillations and wave propagation in a G-protein based model of signal transduction in Dictyostelium discoideum. *Philos. Trans. Roy. Soc. London Ser. B* **349** (1995), 179–195.

[44] Tomchik, K. J., and Devreotes, P. N., Cyclic AMP waves in Dictyostelium discoideum: A demonstration by isotope dilution fluorography. *Science* **212** ( 1981), 443–446.

[45] Velázquez, J. J. L., Stability of some mechanisms of chemotactic aggregation. *SIAM J. Appl. Math.* **62** (5) (2002), 1581–1633.

[46] Velázquez, J. J. L., Point dynamics for a singular limit of the Keller-Segel model I: Motion of the concentration regions. *SIAM J. Appl. Math.* **64** (4) (2004), 1198–1223.

[47] Velázquez, J. J. L., Point dynamics for a singular limit of the Keller-Segel model II: Formation of the concentration regions. *SIAM J. Appl. Math.* **64** (4) (2004), 1224–1248.

[48] Velázquez, J. J. L., Well posedness of a model of point dynamics for a limit of the Keller-Segel problem. *J. Differential Equations* **206** (2) (2004), 315–352.

[49] Velázquez, J. J. L., Curvature blow-up in perturbations of minimal cones evolving by mean curvature flow. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **21** (4) (1994), 595–628.

Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, Madrid 28040, Spain

E-mail: jj_velazquez@mat.ucm.es

# From topological field theory to deformation quantization and reduction

Alberto S. Cattaneo*

**Abstract.** This note describes the functional-integral quantization of two-dimensional topological field theories together with applications to problems in deformation quantization of Poisson manifolds and reduction of certain submanifolds. A brief introduction to smooth graded manifolds and to the Batalin–Vilkovisky formalism is included.

**Mathematics Subject Classification (2000).** Primary 81T45; Secondary 51P05, 53D55, 58A50, 81T70.

**Keywords.** Topological quantum field theory, BV formalism, graded manifolds, deformation quantization, formality, Poisson reduction, $L_\infty$- and $A_\infty$-algebras.

## 1. Introduction: a 2D TFT

**1.1. The basic setting.** Let $\Sigma$ be a smooth compact 2-manifold. On $\mathcal{M}_1 := \Omega^0(\Sigma) \oplus \Omega^1(\Sigma)$ one may define the following very simple action functional:

$$S(\xi, \eta) := \int_\Sigma \eta \, \mathrm{d}\xi, \quad \xi \in \Omega^0(\Sigma), \ \eta \in \Omega^1(\Sigma), \tag{1.1}$$

which is invariant under the distribution $0 \oplus \mathrm{d}\beta$, $\beta \in \Omega^0(\Sigma)$. If we take $\xi$ and $\eta$ as above as coordinates on $\mathcal{M}_1$, we may also write

$$\delta_\beta \xi = 0, \quad \delta_\beta \eta = \mathrm{d}\beta. \tag{1.2}$$

The critical points are closed 0- and 1-forms. As symmetries are given by exact forms, the space of solution modulo symmetries, to which we will refer as the moduli space of solutions, is $H^0(\Sigma) \oplus H^1(\Sigma)$, which is finite dimensional. Moreover, it depends only on the topological type of $\Sigma$. Actually, something more is true: the action of the group of diffeomorphisms connected to the identity is included in the symmetries restricted to the submanifold of critical points. In fact, for every vector field $Y$ on $\Sigma$, we have $\mathrm{L}_Y \xi = \iota_Y \mathrm{d}\xi$ and $\mathrm{L}_Y \eta = \iota_Y \mathrm{d}\eta + \mathrm{d}\iota_Y \eta$. So upon setting $\mathrm{d}\xi = \mathrm{d}\eta = 0$, we get $\mathrm{L}_Y = \delta_{\beta_Y}$ with $\beta_Y = \iota_Y \eta$. This is the simplest example of 2-dimensional topological field theory (TFT) that contains derivatives in the fields.[1]

*The author acknowledges partial support of SNF Grant No. 200020-107444/1.

[1]This example belongs to the larger class of the so-called $BF$ theories. This is actually a 2-dimensional abelian $BF$ theory.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

One may also allow $\Sigma$ to have a boundary $\partial\Sigma$. If we do not impose boundary conditions, the variational problem yields the extra condition i) $\iota_{\partial\Sigma}^*\eta = 0$ where $\iota_{\partial\Sigma}$ denotes the inclusion map of $\partial\Sigma$ into $\Sigma$. So it makes sense to impose i) from the beginning. The second possibility is to impose the boundary condition ii) that $\xi_{|\partial\Sigma}$ should be constant. By translasting $\xi$, we may always assume this constant to be zero.[2] For the symmetries to be consistent with boundary conditions i), we have to assume that $\beta_{|\partial\Sigma}$ is constant, and again we may assume without loss of generality that this constant vanishes. So we consider the following two cases:

| | | |
|---|---|---|
| Neumann boundary conditions: | $\iota_{\partial\Sigma}^*\eta = 0, \quad \beta_{\|\partial\Sigma} = 0$ | (N) |
| Dirichlet boundary conditions: | $\xi_{\|\partial\Sigma} = 0,$ | (D) |

**1.2. Generalizations.** To make things more interesting, we may replicate $n$ times what we have done above. Namely, take $\mathcal{M}_n = \mathcal{M}_1^n$ and define

$$S(\{\xi\},\{\eta\}) := \int_\Sigma \sum_{I=1}^n \eta_I \,\mathrm{d}\xi^I, \quad \xi^I \in \Omega^0(\Sigma),\ \eta_I \in \Omega^1(\Sigma).$$

Identifying $\mathcal{M}_n$ with $\Omega^0(\Sigma,\mathbb{R}^n)\oplus\Omega^1(\Sigma,(\mathbb{R}^n)^*)$, we may also write

$$S(\xi,\eta) := \int_\Sigma \langle \eta, \mathrm{d}\xi\rangle, \quad \xi \in \Omega^0(\Sigma,\mathbb{R}^n),\ \eta \in \Omega^1(\Sigma,(\mathbb{R}^n)^*), \tag{1.3}$$

where $\langle\ ,\ \rangle$ denotes the canonical pairing. The symmetries are now defined by the addition to $\eta$ of an exact 1-form $\mathrm{d}\beta$, $\beta \in \Omega^1(\Sigma,(\mathbb{R}^n)^*)$. If $\Sigma$ has a boundary, we then choose N or D boundary conditions for each index $I$. Accordingly the boundary components of $\beta$ corresponding to N boundary conditions have to be set to zero.

We may also modify the action functional by adding a local term

$$S_\alpha(\xi,\eta) = \frac{1}{2}\int_\Sigma \alpha(\xi)(\eta,\eta), \tag{1.4}$$

where $\alpha$ is a smooth map $\mathbb{R}^n \to \Lambda^2\mathbb{R}^n$ or more generally an element of $\hat{S}(\mathbb{R}^n)^*\otimes\Lambda^2\mathbb{R}^n$, where $\hat{S}(\mathbb{R}^n)$ denotes the formal completion (i.e., the space of formal power series) of the symmetric algebra $S(\mathbb{R}^n)^*$. We will discuss in the following under which assumption on $\alpha$ and on the boundary conditions, this term may be added without breaking the symmetries of $S$.

A further generalization with a smooth $n$-manifold $M$ as target exists. The space $\mathcal{M}(M) := \{\text{bundle maps } T\Sigma \to T^*M\}$ fibers over $\mathrm{Map}(\Sigma, M)$ with fiber at a map $X$ the space of sections $\Gamma(T^*\Sigma\otimes X^*T^*M)$. Regarding $\mathrm{d}X$ as a section of $X^*TM$ and using the canonical pairing $\langle\ ,\ \rangle$ of $TM$ with $T^*M$, we define

$$S(X,\eta) := \int_\Sigma \langle \eta, \mathrm{d}X\rangle, \quad X \in \mathrm{Map}(\Sigma, M),\ \eta \in \Gamma(T^*\Sigma\otimes X^*T^*M). \tag{1.5}$$

[2] For simplicity, in this note we do not consider the case [16] when the boundary is divided into different components with different boundary conditions.

The critical points are now given by pairs of a constant map $X$ and a closed form $\eta \in \Omega^1(\Sigma, T^*_x M)$ with $x = X(\Sigma)$. The symmetries are given by translating $\eta$ by $\mathrm{d}\beta$ with $\beta \in \Gamma(X^*T^*M)$.[3] For the boundary conditions, one chooses a submanifold $C$ of $M$ and imposes $X(\partial\Sigma) \subset C$ and $\iota^*_{\partial\Sigma}\eta \in \Gamma(T^*\partial\Sigma \otimes X^*N^*C)$, where the conormal bundle $N^*C$ is by definition the annihilator of $TC$ as a subbundle of $T_C M$; viz.:

$$N^*_x C := \{\alpha \in T^*_x M : \alpha(v) = 0 \text{ for all } v \in T_x C\}, \quad x \in C. \tag{1.6}$$

Accordingly, we require $\iota^*_{\partial\Sigma}\beta \in \Gamma(X^*N^*C)$. Observe that the tangent space at a given solution (i.e., $X(\Sigma) = x$, $\eta$ closed), is isomorphic – upon choosing local coordinates around $x$ – to $\mathcal{M}_n$, just by setting $X = x + \xi$. Moreover, the action evaluated around a solution is precisely (1.3).

A global generalization of (1.4) is also possible. Namely, to every bivector field $\pi$ (i.e., a section of $\Lambda^2 TM$), we associate the term

$$S_\pi(X, \eta) = \frac{1}{2}\int_\Sigma \pi(X)(\eta, \eta). \tag{1.7}$$

If we work in the neighborhood of a solution $x$ and set $X = x + \xi$, then (1.7) reduces to (1.4) with $\alpha(v) = \pi(x+v), \xi \in \mathbb{R}^n \simeq T_x M$. Actually we are interested in working in a formal neighborhood, so we set $\alpha$ to be the Taylor expansion of $\pi$ around $x$ and regard it as an element of $\hat{S}(\mathbb{R}^n)^* \otimes \Lambda^2\mathbb{R}^n$.

**1.3. Functional-integral quantization.** The action functional (1.5) is not very interesting classically. Much more interesting is its quantization, by which we mean the evaluation of "expectation values", i.e., ratios of functional integrals

$$\langle \mathcal{O} \rangle_{\mathrm{cl}} := \frac{\int_{\mathcal{M}(M)} \mathrm{e}^{\frac{\mathrm{i}}{\hbar}S}\,\mathcal{O}}{\int_{\mathcal{M}(M)} \mathrm{e}^{\frac{\mathrm{i}}{\hbar}S}}, \tag{1.8}$$

where $\mathcal{O}$ is a function (which we assume to be a polynomial or a formal power series) on $\mathcal{M}(M)$. The evaluation of these functional integrals consists of an ordinary integration over the moduli space of solutions and of an "infinite-dimensional integral" which is operatively defined in terms of the momenta of the Gaussian distribution given by $S$.

The finite-dimensional integration is not problematic, though it requires choosing a measure on the moduli space of solution. The main assumption in this paper is that the first cohomology of $\Sigma$ with whatsoever boundary conditions is trivial. Actually, we assume throughout that $\Sigma$ is the 2-disk $D$. Up to equivalence then a solution is given by specifying the value $x$ of the constant map $X$. Thus, the moduli space of solution is $M$. We choose then a delta measure on $M$ at some point $x$.

The second integration, performed around a point $x$, is then over $\mathcal{M}_n$. The main problem is that the operator d defining the quadratic form in $S$ is not invertible. To

[3] The derivative of $\beta$ is computed by choosing any torsion-free connection on $M$.

overcome this problem and make sense of the integration, we resort to the so-called BV (Batalin–Vilkovisky [5]) formalism, which is reviewed in Section 3. Besides giving us an operative unambiguous definition of (1.8), the BV formalism will also provide us with relations among the expectation values, the so-called Ward identities (see 4.5). The latter computation is however less rigorous; one might think of this as a machinery suggesting relations that have next to be proven to hold. Moreover, the BV formalism leads naturally to the generalization when the target $M$ is a graded manifold (see Section 2). In this context there is an interesting duality (see 4.3 and 4.4) between different targets.

**Acknowledgment.** I thank F. Bonechi, D. Fiorenza, F. Helein, R. Mehta, C. Rossi, F. Schätz, J. Stasheff and M. Zambon for very useful comments.

## 2. Smooth graded manifolds

In this section we give a crash course in the theory of smooth graded manifolds. A graded manifold is a supermanifold with a $\mathbb{Z}$-refinement of the $\mathbb{Z}_2$-grading. As we work in the smooth setting, we can work with algebras of global functions and so avoid the more technical definitions in terms of ringed spaces. We begin with recalling some basic definitions and notations.

**2.1. Graded linear algebra.** A graded vector space $V$ is a direct sum over $\mathbb{Z}$ of vector spaces: $V = \oplus_{i\in\mathbb{Z}} V_i$. Elements of $V_i$ have by definition degree $i$. By $V[n]$, $n \in \mathbb{Z}$, we denote the graded vector space with the same components of $V$ but shifted by $n$; i.e., $V[n]_i := V_{i+n}$. A morphism $\phi\colon V \to W$ of graded vector spaces is a homomorphism that preserves degree: i.e., $\phi(V_i) \subset W_i$ for all $i$. A $j$-graded homomorphism $\phi\colon V \to W$ is a morphism $V \to W[j]$; i.e., $\phi(V_i) \subset W_{i+j}$. We denote by $\mathrm{Hom}_j(V, W)$ the space of $j$-graded homomorphisms. We may regard the vector space of homomorphisms as a graded vector space $\mathrm{Hom}(V, W) = \oplus_j \mathrm{Hom}_j(V, W)$. In particular, by regarding the ground field as a graded vector space concentrated in degree zero, the dual $V^*$ of a graded vector space $V$ is also naturally graded with $V_i^* := (V^*)_i$ isomorphic to $(V_{-i})^*$. Observe that $V[n]^* = V^*[-n]$. Tensor products of graded vector spaces are also naturally graded: $(V \otimes W)_i = \oplus_{r+s=i} V_r \otimes W_s$.

**2.1.1. Graded algebras.** A graded algebra $A$ is an algebra which is also a graded vector space such that the product is a morphism of graded vector spaces. The algebra is called graded commutative (skew-commutative) if $ab = (-1)^{ij}ba$ ($ab = -(-1)^{ij}ba$) for all $a \in A_i$, $b \in A_j$, $i, j \in \mathbb{Z}$. The symmetric algebra of a graded vector space is the graded commutative algebra defined as $S(V) = T(V)/I$, where $T(V)$ denotes the tensor algebra and $I$ is the two-sided ideal generated by $vw - (-1)^{ij}wv$, $v \in V_i$, $w \in V_j$. We denote by $\hat{S}(V)$ its formal completion consisting of formal power series.

A graded skew-commutative algebra is called a graded Lie algebra (GLA) if its product, denoted by $[\ ,\ ]$ satisfies the graded Jacobi identity: $[a,[b,c]] = [[a,b],c] + (-1)^{ij}[b,[a,c]]$, for all $a \in A_i$, $b \in A_j$, $c \in A$, $i, j \in \mathbb{Z}$.

**2.1.2. Graded modules.** A graded module $M$ over a graded algebra $A$ is a graded vector space which is a module over $A$ regarded as a ring such that the action $A \otimes M \to M$ is a morphism of graded vector spaces. If $M$ is a module, then so is $M[j]$ for all $j \in \mathbb{Z}$.

The tensor product $M_1 \otimes_A M_2$ over $A$ of a right $A$-module $M_1$ and a left $A$-module $M_2$ is defined as the quotient of $M_1 \otimes M_2$ by the subspace generated by $m_1 a \otimes m_2 - m_1 \otimes a m_2$, for all $a_{\in} A$, $m_i \in M_i$. Observe that if $M_1$ and $M_2$ are bimodules, then so is $M_1 \otimes_A M_2$.

Let $M$ be a left $A$-module. If $A$ is graded commutative (skew-commutative), we make $M$ into a bimodule by setting $ma := (-1)^{ij} am$ ($ma := -(-1)^{ij} am$) , $a \in A_i$, $m \in M_j$. We may regard $A \oplus M$ as a graded commutative (skew-commutative) algebra by setting the product of two elements in $M$ to zero. If $A$ is a GLA, then so is $A \oplus M$.

Let $A$ be graded commutative. For every $A$-module $M$, we define inductively the $A$-module $T_A^k(M)$ as $T_A^{k-1}(M) \otimes_A M$, with $T_A^0(M) := A$. So one gets the graded associative algebra $T_A(M) := \oplus_{j\in\mathbb{N}} T_A^j(M)$ which is also an $A$-bimodule. The symmetric algebra $S_A(M)$ is defined as the quotient of $T_A(M)$ by the two-sided ideal generated by $vw - (-1)^{ij} wv$, $v \in M_i$, $w \in M_j$. We denote by $\hat{S}_A(M)$ its formal completion.

**2.1.3. Derivations and multiderivations.** A $j$-graded endomorphism $D$ of a graded algebra $A$ is called a $j$-graded derivation if $D(ab) = D(a)b + (-1)^{ij} aD(b)$ for all $a \in A_i$, $i \in \mathbb{Z}$, and all $b \in A$. For example, if $A$ is a GLA, $[a,\ ]$ is an $i$-graded derivation for every $a \in A_i$. A differential is a derivation of degree 1 that squares to zero. A differential graded Lie algebra (DGLA) is a GLA with a differential.

We denote by $\mathrm{Der}_j(A)$ the space of $j$-graded derivations of a graded algebra $A$ and set $\mathrm{Der}(A) = \oplus_{j\in\mathbb{Z}} \mathrm{Der}_j(A)$. It is a GLA with bracket $[D_1, D_2] := D_1 D_2 - (-1)^{j_1 j_2} D_2 D_1$, $D_i \in \mathrm{Der}_{j_i}(A)$. Observe that $\mathrm{Der}(A)$ is a left $A$-module while $A$ is a left $\mathrm{Der}(A)$-module. Thus, for every $n$, we may regard $\mathrm{Der}(A) \oplus A[n]$ as a GLA with the property

$$\begin{aligned} &[X, fg] = (-1)^{jk} f[X, g] + [X, f]g, \\ &\text{for all } X \in \mathrm{Der}(A)_j,\ f \in A_k,\ g \in A. \end{aligned} \tag{2.1}$$

Given a graded commutative algebra $A$, we define the algebra $\hat{\mathrm{D}}(A, n)$ of $n$-shifted multiderivations by $\hat{\mathrm{D}}(A, n) := \hat{S}_A(\mathrm{Der}(A)[-n])$, and denote by $\mathrm{D}(A, n)$ its subalgebra $S_A(\mathrm{Der}(A)[-n])$. Observe that the GLA structure on $\mathrm{Der}(A) \oplus A[n]$ can be extended to $\mathrm{D}(A, n)[n]$ and to $\hat{\mathrm{D}}(A, n)[n]$ in a unique way, compatible with (2.1), such that

$$[D_1, D_2 D_3] = (-1)^{(j_1+n)j_2} D_2 [D_1, D_3] + [D_1, D_2] D_3, \quad D_i \in \mathrm{D}(A)_{j_i}.$$

By this property, $\hat{\mathrm{D}}(A,n)$ is a so-called $n$-Poisson algebra. For $n=0$, it is a graded Poisson algebra. A 1-Poisson algebra is also called a Gerstenhaber algebra. Since this case is particularly important, we will use the special notation $\hat{\mathrm{D}}(A)$ ($\mathrm{D}(A)$) for $\hat{\mathrm{D}}(A,1)$ ($\mathrm{D}(A,1)$). Elements of $\hat{\mathrm{D}}(A)$ are simply called multiderivations. More precisely, elements of $S^j_A(\mathrm{Der}(A)[-1])$ are called $j$-derivations, and a $j$-derivation is said to be of degree $k$ and of total degree $j+k$ if it belongs to $\hat{\mathrm{D}}(A)_{j+k}$.

Given an $n$-Poisson algebra $(P,\bullet,[\ ,\ ])$, one defines $\mathrm{ad}\colon P\to\mathrm{Der}(P)$ by $\mathrm{ad}_X Y := [X,Y]$, $X,Y\in P$. The $n$-Poisson algebra is said to be nondegenerate if ad is surjective (in other words, if the first Lie algebra cohomology of $P$ with coefficients in its adjoint representation is trivial).

**2.1.4. The Hochschild complex.** For a given a graded vector space $A$ one defines $\mathsf{Hoch}^{j,m}(A)=\mathrm{Hom}_j(A^{\otimes m},A)$, $\mathsf{Hoch}^n(A)=\bigoplus_{j+m=n}\mathsf{Hoch}^{j,m}(A)$, and the Hochschild complex $\mathsf{Hoch}(A)=\bigoplus_n\mathsf{Hoch}^n(A)$. One may compose elements of $\mathsf{Hoch}(A)$ as follows: given $\phi\in\mathsf{Hoch}^{j_1,m_1}$ and $\psi\in\mathsf{Hoch}^{j_2,m_2}$, one defines the nonassociative product

$$\phi\bullet\psi=(-1)^{(j_2+m_2-1)(m_1-1)}\sum_i(-1)^{i(m_2-1)}\phi\circ(1^{\otimes i}\otimes\psi\otimes 1^{\otimes(m_1-1-i)})$$
$$\in\mathsf{Hoch}^{j_1+j_2,m_1+m_2-1}.$$

It turns out that its associated bracket $[\phi,\psi]:=\phi\bullet\psi-(-1)^{(j_1+m_1-1)(j_2+m_2-1)}\psi\bullet\phi$ makes $\mathsf{Hoch}(A)[1]$ into a GLA. A product on $A$ is an element $\mu$ of $\mathsf{Hoch}^{0,2}(A)$. Define $b=[\mu,\ ]$. Then $b$ is a differential on $\mathsf{Hoch}(A)[1]$ iff the product is associative.

**2.1.5. Differential and multidifferential operators.** Given a graded associative algebra $A$ and graded derivations $\phi_i\in\mathrm{Der}(A)_{j_i}$, the composition $\phi_1\circ\cdots\circ\phi_k$ is an element of $\mathsf{Hoch}^{j_1+\cdots+j_k,1}$. A differential operator on $A$ is by definition a linear combination of homomorphisms of this form. A multidifferential operator is a linear combination of elements of $\mathsf{Hoch}(A)$ of the form $(a_1,\dots,a_n)\mapsto\phi_1(a_1)\dots\phi_n(a_n)$ where each $\phi_i$ is a differential operator. Denote by $\mathcal{D}(A)$ the Lie subalgebra of multidifferential operators in $\mathsf{Hoch}(A)[1]$. As the product is a multidifferential operator itself, $\mathcal{D}(A)$ is also a subcomplex of $(\mathsf{Hoch}(A)[1],b)$. For $A$ graded commutative, one defines the HKR map (Hochschild–Kostant–Rosenberg [26]) $\mathrm{HKR}\colon\mathrm{D}(A)\to\mathcal{D}(A)$ as the linear extension of

$$\phi_1\dots\phi_n\mapsto\Big(a_1\otimes\cdots\otimes a_n\mapsto\sum_{\sigma\in S_n}\mathrm{sign}(\sigma)\phi_{\sigma(1)}(a_1)\dots\phi_{\sigma(n)}(a_n)\Big),$$

where the $\phi_i$s are derivations and the sign is given by $\phi_{\sigma(1)}\dots\phi_{\sigma(n)}=\mathrm{sign}(\sigma)\phi_1\dots\phi_n$ in $\mathrm{D}(A)$. It turns out that HKR is a chain map $(\mathrm{D}(A),0)\to(\mathcal{D}(A),b)$. It is a classical result [26] that in certain cases (e.g., when $A$ is the algebra of smooth functions on a smooth manifold), HKR is a quasiisomorphism (i.e., it induces an isomorphism in cohomology).

**2.2. Graded vector spaces.** To fix notations, from now on we assume the ground field to be $\mathbb{R}$. For simplicity we consider only finite-dimensional vector spaces. We define the algebra of polynomial functions over a graded vector space $V$ as the symmetric algebra of $V^*$ and the algebra of smooth functions as its formal completion. We use the notations $\mathbf{C}^\infty(V) := S(V^*) \subseteq \hat{\mathbf{C}}^\infty(V) := \hat{S}(V^*)$. Elements of $S^0(V^*) \simeq \mathbb{R}$ will be called constants functions.

**2.2.1. Multivector fields.** A vector field on $V$ is by definition a linear combination of graded derivations on its algebra of functions. We use the notations $\mathfrak{X}(V) := \mathrm{Der}(\mathbf{C}^\infty(V))$, $\hat{\mathfrak{X}}(V) := \mathrm{Der}(\hat{\mathbf{C}}^\infty(V))$. Observe that we may identify $\mathfrak{X}(V)$ and $\hat{\mathfrak{X}}(V)$ with $\mathbf{C}^\infty(V) \otimes V$ and $\hat{\mathbf{C}}^\infty(V) \otimes V$, respectively. Elements of $S^0(V^*) \otimes V \simeq V$ will be called constant vector fields.

Multivector fields are by definition multiderivations. In particular, $k$-vector fields are $k$-multiderivations, and we define their degree and total degree correspondingly. We use the notations $\boldsymbol{\mathfrak{X}}(V) := \mathrm{D}(\mathbf{C}^\infty(V))$ and $\hat{\boldsymbol{\mathfrak{X}}}(V) := \hat{\mathrm{D}}(\hat{\mathbf{C}}^\infty(V))$ for the corresponding Gerstenhaber algebras. We also define the $n$-Poisson algebras $\boldsymbol{\mathfrak{X}}(V, n)$ and $\hat{\boldsymbol{\mathfrak{X}}}(V, n)$ of $n$-shifted multivector fields as $\mathrm{D}(\mathbf{C}^\infty(V), n)$ and $\hat{\mathrm{D}}(\hat{\mathbf{C}}^\infty(V), n)$. We have the following identifications:

$$\boldsymbol{\mathfrak{X}}(V, n) \simeq S(V^*) \otimes S(V[-n]) \simeq \mathbf{C}^\infty(V \oplus V^*[n]), \tag{2.2a}$$

$$\hat{\boldsymbol{\mathfrak{X}}}(V, n) \simeq \hat{S}(V^*) \hat{\otimes} \hat{S}(V[-n]) \simeq \hat{\mathbf{C}}^\infty(V \oplus V^*[n]). \tag{2.2b}$$

**2.2.2. Berezinian integration.** Let $V$ be an odd vector space (i.e., a graded vector space with nontrivial components only in odd degrees). By integration we simply mean a linear form on its space of functions $\mathbf{C}^\infty(V) = \hat{\mathbf{C}}^\infty(V)$, which is isomorphic, forgetting degrees, to $\Lambda V^*$.[4] So integration is defined by an element $\mu$ of $\Lambda V$. We use the notation $\int_V f\,\mu$ for the pairing $\langle f, \mu\rangle$. We call an element of $\Lambda V$ a Berezinian form if its component in $\Lambda^{\mathrm{top}} V$, top $= \dim V$, does not vanish. In this case integration has the property that its restriction to the space of functions of top degree is injective. A Berezinian form concentrated in top degree, i.e., an element of $\Lambda^{\mathrm{top}} V \setminus \{0\}$, is called pure and has the additional property that the corresponding integral vanishes on functions that are not of top degree. Observe that a pure Berezinian form $\rho$ establishes an isomorphism $\phi_\rho \colon \mathbf{C}^\infty(V) \simeq \Lambda V^* \xrightarrow{\sim} \Lambda V$, $g \mapsto \iota_g \rho$. If $\mu = \iota_g \rho$, then $\int_V f\,\mu = \langle f, \iota_g \rho\rangle = \int_V fg\,\rho$, so we simply write $g\rho$ instead of $\iota_g \rho$.

**Lemma 2.1.** *Given a pure Berezinian form $\rho$, for every Berezinian form $\mu$ there is a unique constant $c \neq 0$ and a unique function $\sigma \in \Lambda^{>0} V^*$ such that $\mu = c\mathrm{e}^\sigma \rho$.*

*Proof.* Set $g = \phi_\rho^{-1}(\mu)$. If $\mu$ is a Berezinian form, its component $c$ in $\Lambda^0 V^*$ is invertible. So we may write, $g = c(1 + h)$ with $h \in \Lambda^{>0} V^*$. Finally we define $\sigma = \log(1+h) = \sum_{k=1}^\infty (-1)^{k+1} h^k / k$ (observe that this is actually a finite sum). □

[4] By $\Lambda V$, we mean the usual exterior algebra of $V$ regarded as an ordinary vector space.

**Lemma 2.2.** *For every Berezinian form $\mu$, there is a map* $\mathrm{div}_\mu\colon \mathfrak{X}(V) \to \mathbf{C}^\infty(V)$ (*the divergence operator*) *such that*

$$\int_V X(f)\,\mu = \int_V f\,\mathrm{div}_\mu X\,\mu \quad \textit{for all } f \in \mathbf{C}^\infty(V).$$

*Moreover,* $\mathrm{div}_{c\mu} = \mathrm{div}_\mu$ *for every constant* $c \neq 0$. *In particular, all pure Berezinian forms define the same divergence operator.*

*Proof.* The map $f \mapsto \int_V X(f)\,\mu$ is linear. So there is a unique $\mu_X \in \Lambda V$ such that $\int_V X(f)\,\mu = \int_V f\,\mu_X$. Given a pure Berezinian form $\rho$, define $g_\mu = \phi_\rho^{-1}(\mu)$ and $g_X^\mu = \phi_\rho^{-1}(\mu_X)$. Thus, $\mu_X = g_X^\mu \rho = g_X^\mu g_\mu^{-1}\mu$. Then we define $\mathrm{div}_\mu X$ as $g_X^\mu g_\mu^{-1}\mu$. Observe that this does not depend on the choice of $\rho$. □

**2.3. Graded vector bundles.** A graded vector bundle is a vector bundle whose fibers are graded vector spaces and such that the transition functions are morphisms of graded vector spaces. All the constructions for graded vector spaces described above extend to graded vector bundles. In particular, given a graded vector bundle $E$, we may define the shifted graded vector bundles $E[n]$, the dual bundle $E^*$ (and $E[n]^* = E^*[-n]$), the symmetric algebra bundle $S(E)$ and its formal completion $\hat{S}(E)$. We also define the graded commutative algebras of functions (we restrict for simplicity to graded vector bundles of finite rank) accordingly in terms of sections $\mathbf{C}^\infty(E) := \Gamma(S(E^*)) \subseteq \hat{\mathbf{C}}^\infty(E) := \Gamma(\hat{S}(E^*))$. Elements of $\mathbf{C}^\infty(E)$ will be called polynomial functions.

**Remark 2.3.** In case the given vector bundle is the tangent or the cotangent bundle of a manifold $M$, it is customary to write the shift after the $T$ symbol; viz., one writes $T[n]M$ and $T^*[n]M$ instead of $TM[n]$ and $T^*M[n]$. According to the previous remark, we have in particular $\mathbf{C}^\infty(T[1]M) = \hat{\mathbf{C}}^\infty(T[1]M) = \Omega(M)$ and $\mathbf{C}^\infty(T^*[1]M) = \hat{\mathbf{C}}^\infty(T^*[1]M) = \mathcal{X}(M)$, where $\Omega(M) = \Gamma(\Lambda T^*M)$ and $\mathcal{X}(M) = \Gamma(\Lambda TM)$ denote the graded commutative algebras of differential forms and of multivector fields respectively. Observe that, in terms of graded vector spaces, we have

$$\Omega(M) = \bigoplus_{i=0}^{\dim M} \Omega^i(M)[-i], \quad \mathcal{X}(M) = \bigoplus_{i=0}^{\dim M} \mathcal{X}^i(M)[-i], \tag{2.3}$$

where $\Omega^i(M)$ and $\mathcal{X}^i(M)$ are regarded as ordinary vector spaces.

**2.3.1. Multivector fields.** A vector field on $E$ is a linear combination of graded derivations on its algebra of functions. We use the notations $\mathfrak{X}(E) := \mathrm{Der}(\mathbf{C}^\infty(E))$, $\hat{\mathfrak{X}}(E) := \mathrm{Der}(\hat{\mathbf{C}}^\infty(E))$. A vector field $X$ on $E$ is completely determined by its restrictions $X_M$ to $C^\infty(M)$ and $X_E$ to $\Gamma(E^*)$. Observe that $X_M$ is a $\hat{\mathbf{C}}^\infty(E)$-valued vector field on $M$. Picking a connection $\nabla$ on $E^*$, we set $X_E^\nabla(\sigma) := X(\sigma) - \nabla_{X_M}\sigma$,

for all $\sigma \in \Gamma(E^*)$. Since $X_E^\nabla$ is $C^\infty(M)$-linear, it defines a bundle map $E^* \to \hat{S}(E^*)$. The map $X \mapsto X_M \oplus X_E^\nabla$ is then an isomorphism from $\hat{\mathfrak{X}}(E)$ to $\Gamma(\hat{S}E^* \otimes (TM \oplus E))$.

**Remark 2.4.** We may extend $\nabla$ to the whole of $\hat{\mathbf{C}}^\infty(E)$ as a derivation. So $\nabla_{X_M}$, unlike $X_M$, is a vector field on $E$. The difference $X^\nabla := X - \nabla_{X_M}$, which we call the vertical component of $X$, is then also a vector field with the additional property that its restriction to $C^\infty(M)$ vanishes.

Multivector fields are by definition multiderivations. In particular, $k$-vector fields are $k$-multiderivations, and we define their degree and total degree correspondingly. By $\mathfrak{X}(E) := \mathrm{D}(\mathbf{C}^\infty(E))$, $\hat{\mathfrak{X}}(E) := \hat{\mathrm{D}}(\hat{\mathbf{C}}^\infty(E))$ we denote the corresponding Gerstenhaber algebras. More generally, the $n$-Poisson algebra $\hat{\mathfrak{X}}(E,n)$ ($\mathfrak{X}(E,n)$) of $n$-shifted (polynomial) multivector fields are defined as $\hat{\mathrm{D}}(\hat{\mathbf{C}}^\infty(E),n)$ ($\mathrm{D}(\mathbf{C}^\infty(E),n)$). Upon choosing a connection $\nabla$, we have the identifications

$$\mathfrak{X}(E,n) \simeq \Gamma(SE^*) \otimes \Gamma(S((TM \oplus E)[-n])) \simeq \mathbf{C}^\infty(E \oplus T^*[n]M \oplus E^*[n]),$$
$$\hat{\mathfrak{X}}(E,n) \simeq \Gamma(\hat{S}(E^*)) \hat{\otimes} \Gamma(\hat{S}(TM \oplus E)[-n]) \simeq \hat{\mathbf{C}}^\infty(E \oplus T^*[n]M \oplus E^*[n]).$$

**2.3.2. The Berezinian bundle.** We may easily extend the Berezinian integration introduced in 2.2.2 to every odd vector bundle $E \to M$ (i.e., a bundle of odd vector spaces). A section $\mu$ of the "Berezinian bundle" $\mathrm{BER}(E) := \Lambda E \otimes \Lambda^{\mathrm{top}} T^*M$, $\mathrm{top} = \dim M$, defines[5] a $C^\infty(M)$-linear map $\langle\ ,\mu\rangle\colon \mathbf{C}^\infty(E) \simeq \Gamma(\Lambda E^*) \to \Omega^{\mathrm{top}}(M)$. We set $\int_E f\,\mu := \int_M \langle f,\mu\rangle$. (For $M$ non compact, this of course makes sense only for certain functions.) Like in the case of odd vector spaces, we are interested in integrations that are nondegenerate on the subspace of functions of top degree. These are determined by sections of the Berezinian bundle whose top component is nowhere vanishing. We call such sections Berezinian forms. A pure Berezinian form $\rho$ is then by definition a Berezinian form concentrated in top degree, i.e., a nowhere vanishing section of the "pure Berezinian bundle" $\mathrm{Ber}(E) := \Lambda^{\mathrm{top}} E \otimes \Lambda^{\mathrm{top}} T^*M$ (with the first "top" the rank of $E$).

**Example 2.5.** Let $E = T^*[k]M$, with $k$ odd and with $M$ orientable and connected. Then $\mathrm{Ber}(E) = (\Lambda^{\mathrm{top}} T^*M)^{\otimes 2}$. So there is a two-to-one correspondence between volume forms on $M$ and pure Berezinian forms on $E$. Let $v$ be a volume form and $\rho_v$ the corresponding Berezinian form. If we identify functions on $T^*[k]M$ with multivector fields, we may then compute $\int_{T^*[k]M} X\,\rho_v = \int_M \phi_v(X)\,v$, with $\phi_v\colon \mathfrak{X}(M) \xrightarrow{\sim} \Omega(M)$, $X \mapsto \iota_X v$. As a further example, consider the graded vector bundle $L_C := N^*[k]C, k$ odd, where $C$ is a submanifold of $M$ and $N^*C$ its conormal bundle (defined in (1.6)). Now $\mathrm{Ber}\, L_C \simeq \Lambda^{\mathrm{top}} N^*C \otimes \Lambda^{\mathrm{top}} T^*C \simeq \Lambda^{\mathrm{top}} T^*_C M$, where $T^*_C M$ is the restriction of $T^*M$ to $C$. Thus, a volume form $v$ on $M$ also determines by restriction a pure Berezinian form on $L_C$ which we denote by $\sqrt{\rho_v}$ as the correspondence is now linear instead of quadratic. We may identify functions on $L_C$ with sections of the exterior

[5] We consider $M$ to be orientable, otherwise replace the space of top forms with the space of densities.

algebra of $NC$. We then have $\int_{L_C} X \sqrt{\rho_v} = \int_C \phi_v(\tilde{X})$, where $\tilde{X}$ is any multivector field on $M$ extending a representative of $X$ in $\Gamma(\Lambda T_C M)$. Finally, we have a canonically defined surjective morphism $\imath_C^* \colon \mathbf{C}^\infty(M) \to \mathbf{C}^\infty(C)$ obtained by restricting a multivector field to $C$ and modding out its tangent components. One should think of $L_C$ as a submanifold (actually, a Lagrangian submanifold) of $T^*[k]M$ with inclusion map denoted by $\imath_C$. We then have

$$\int_{L_c} \imath_C^*(X) \sqrt{\rho_v} = \int_C \phi_v(X) \quad \text{for all } X \in \Gamma(\Lambda T M) \simeq \mathbf{C}^\infty(T^*[k]M), \tag{2.4}$$

with the r.h.s. defined to be zero if form degree and dimension do not match.

A pure Berezinian form $\rho$ establishes an isomorphism $\phi_\rho \colon \mathbf{C}^\infty(E) \simeq \Gamma(\Lambda E^*) \xrightarrow{\sim} \Gamma(\mathrm{BER}(E))$, $g \mapsto \imath_g \rho$. If $\mu = \imath_g \rho$, then $\int_E f\,\mu = \int_M \langle f, \imath_g \rho \rangle = \int_E fg\,\rho$, so we simply write $g\rho$ instead of $\imath_g \rho$. Lemmata 2.1 and 2.2 generalize as follows:

**Lemma 2.6.** *Given a pure Berezinian form $\rho$, for every Berezinian form $\mu$ there is a unique nowhere vanishing function $f \in C^\infty(M)$ and a unique function $\sigma \in \Gamma(\Lambda^{>0} E^*)$ such that $\mu = f \mathrm{e}^\sigma \rho$. If $M$ is connected, there is a unique function $\sigma \in \mathbf{C}^\infty(E)$ such that $\mu = \mathrm{e}^\sigma \rho$ or $\mu = -\mathrm{e}^\sigma \rho$.*

**Lemma 2.7.** *Let $E \to M$ be an odd vector bundle with $M$ compact and orientable. Then, for every Berezinian form $\mu$, there is a map $\mathrm{div}_\mu \colon \mathfrak{X}(E) \to \mathbf{C}^\infty(E)$ (the divergence operator) such that*

$$\int_E X(f)\,\mu = \int_E f\,\mathrm{div}_\mu X\,\mu \quad \textit{for all } f \in \mathbf{C}^\infty(E).$$

*Moreover,* $\mathrm{div}_{c\mu} = \mathrm{div}_\mu$ *for every constant* $c \neq 0$.

The proof of Lemma 2.6 is exactly the same as the proof of Lemma 2.1. The proof of Lemma 2.7 goes as the proof of Lemma 2.2 if we may assume that the map $f \mapsto \langle X(f), \mu \rangle$ is $C^\infty(M)$-linear. This is the case only for a vertical vector field. By using Remark 2.4, we write $X$ as $\nabla_{X_M} + X^\nabla$, and $X^\nabla$ is vertical. By further writing $X_M$ as $\sum_i h_i X_M^i$, with $h_i \in \mathbf{C}^\infty(E)$ and $X_M^i \in \mathfrak{X}(M)$, and manipulating the integral carefully, we end up with terms which are $C^\infty(M)$-linear plus terms where we may apply the usual divergence theorem on $M$. The expression for $\mathrm{div}_\mu X$ is then easily seen not to depend on the choices involved in this argument.

**Remark 2.8.** One may easily see that for every vector field $X$ and every function $g$, the divergence of $gX$ is the sum (with signs) of $g\mathrm{div}_\mu X$ and $X(g)$.

Integration over an arbitrary graded vector bundle is defined by splitting it into its odd part (where Berezinian integration may be defined) and its even part (where the usual integration theory makes sense).

**2.4. Smooth graded manifolds.** We are now ready to define smooth graded manifolds. We call a graded commutative algebra a graded algebra of smooth (polynomial) functions if it is isomorphic to the algebra of (polynomial) functions of a graded vector bundle. Next we denote by $\widehat{\mathsf{GrSmFun}}$ ($\mathsf{GrSmFun}$) the category whose objects are graded algebras of smooth (polynomial) functions and whose morphisms are graded algebra morphisms. Finally, we define the category $\widehat{\mathsf{SmoothGr}}$ ($\mathsf{SmoothGr}$) of smooth graded manifolds as the dual of $\widehat{\mathsf{GrSmFun}}$ ($\mathsf{GrSmFun}$). In particular, graded vector spaces and graded vector bundles may be regarded as smooth graded manifolds, i.e., as objects in $\widehat{\mathsf{SmoothGr}}$ or $\mathsf{SmoothGr}$ depending on which algebra of functions we associate to them.

**Notation 2.9.** If $A$ is an object of $\mathsf{GrSmFun}$, we write $\mathrm{Spec}(A)$ for the same object in $\mathsf{SmoothGr}$. Vice versa, if we start with an object $\mathcal{M}$ of $\mathsf{SmoothGr}$, we denote by $\mathbf{C}^\infty(\mathcal{M})$ the same object in $\mathsf{GrSmFun}$. We use the notations $\widehat{\mathrm{Spec}}$ and $\hat{\mathbf{C}}^\infty$ for the hatted categories. We denote by $\widehat{\mathrm{Mor}}(\mathcal{M}, \mathcal{N})$ ($\mathrm{Mor}(\mathcal{M}, \mathcal{N})$) the space of morphisms from $\mathcal{M}$ to $\mathcal{N}$ in $\widehat{\mathsf{SmoothGr}}$ ($\mathsf{SmoothGr}$).

**Remark 2.10.** The spaces of morphisms $\widehat{\mathrm{Mor}}(\mathcal{M}, \mathcal{N})$ ($\mathrm{Mor}(\mathcal{M}, \mathcal{N})$) may actually be given the structure of a (possibly infinite-dimensional) smooth manifolds. In particular, for $\mathcal{N} = V$ a graded vector space, they may be regarded as (possibly infinite-dimensional) vector spaces:

$$\mathrm{Mor}(\mathcal{M}, V) \simeq (V \otimes \mathbf{C}^\infty(\mathcal{M}))_0, \quad \widehat{\mathrm{Mor}}(\mathcal{M}, V) \simeq (V \otimes \hat{\mathbf{C}}^\infty(\mathcal{M}))_0, \tag{2.5}$$

for $\mathbf{C}^\infty(V)$ is generated by $V^*$, so an algebra morphism from $\mathbf{C}^\infty(V)$ is determined by its restriction to $V^*$ as a morphism of graded vector spaces.

By our definition, every smooth graded manifold may actually be realized as a graded vector bundle though not in a canonical way. One often obtains new graded algebras of smooth functions by some canonical constructions, yet their realization as algebras of functions of graded vector bundles requires some choice.

**Example 2.11.** As we have seen at the end of 2.3.1, upon choosing a connection, we may identify the algebra $\hat{\mathfrak{X}}(E, n)$ of shifted multivector fields on $E$ with the graded algebra of smooth functions on $E \oplus T^*[n]M \oplus E^*[n]$. We write $T^*[n]E$ for $\mathrm{Spec}\, \hat{\mathfrak{X}}(E, n)$ and have, tautologically, $\hat{\mathbf{C}}^\infty(T^*[n]E) = \hat{\mathfrak{X}}(E, n)$ and, noncanonically, $T^*[n]E \simeq E \oplus T^*[n]M \oplus E^*[n]$.

Given two smooth graded manifolds $\mathcal{M}$ and $\mathcal{N}$, their Cartesian product $\mathcal{M} \times \mathcal{N}$ is defined as the smooth graded manifold having $\mathbf{C}^\infty(\mathcal{M}) \hat{\otimes} \mathbf{C}^\infty(\mathcal{N})$ as algebra of functions (respectively $\hat{\mathbf{C}}^\infty(\mathcal{M}) \hat{\otimes} \hat{\mathbf{C}}^\infty(\mathcal{N})$ in the hatted category).

**Remark 2.12** (Graded maps). Unlike in the category of manifolds, in general $\mathrm{Mor}(\mathcal{L} \times \mathcal{M}, \mathcal{N})$ is not the same as $\mathrm{Mor}(\mathcal{L}, \mathrm{Mor}(\mathcal{M}, \mathcal{N}))$ even allowing infinite-dimensional objects. However, one can show that, given $\mathcal{M}$ and $\mathcal{N}$, the functor defined by $\mathcal{L} \mapsto \mathrm{Mor}(\mathcal{L} \times \mathcal{M}, \mathcal{N})$ is representable by an infinite-dimensional

smooth graded manifold [44], [36] denoted by $\mathrm{Map}(\mathcal{M}, \mathcal{N})$; viz., $\mathrm{Mor}(\mathcal{L}\times\mathcal{M}, \mathcal{N}) = \mathrm{Mor}(\mathcal{L}, \mathrm{Map}(\mathcal{M}, \mathcal{N}))$. Similarly, there is a hatted version denoted by $\widehat{\mathrm{Map}}(\mathcal{M}, \mathcal{N}))$.

For $\mathcal{N} = V$ a graded vector space, one can use (2.5)[6] and realize the graded manifolds of maps as graded vector spaces. Namely, one can easily show that

$$\mathrm{Map}(\mathcal{M}, V) \simeq V \otimes \mathbf{C}^\infty(\mathcal{M}), \quad \widehat{\mathrm{Map}}(\mathcal{M}, V) \simeq V \otimes \hat{\mathbf{C}}^\infty(\mathcal{M}). \tag{2.6}$$

In particular, one has the useful identities $\mathbf{C}^\infty(\mathcal{M}) \simeq \mathrm{Map}(\mathcal{M}, \mathbb{R})$, $\mathrm{Mor}(\mathcal{M}, V) = \mathrm{Map}(\mathcal{M}, V)_0, \mathrm{Map}(\mathcal{M}, V[k]) = \mathrm{Map}(\mathcal{M}, V)[k], \mathrm{Map}(\mathcal{M}, V\oplus W) = \mathrm{Map}(\mathcal{M}, V)\oplus \mathrm{Map}(\mathcal{M}, W)$, and their hatted versions.

On a graded manifold we can then define the notions of vector fields, multivector fields, Berezinian integration, divergence operator. In particular, if $\mathcal{M}$ is a smooth graded manifold with algebra of functions isomorphic to $\hat{\mathbf{C}}^\infty(E)$ for some graded vector bundle $E$, we have that $\hat{\boldsymbol{\mathfrak{X}}}(\mathcal{M}, n) := \hat{\mathrm{D}}(\hat{\mathbf{C}}^\infty(\mathcal{M}), n)$ is isomorphic to $\hat{\boldsymbol{\mathfrak{X}}}(E, n)$, so it is a graded algebra of smooth functions. We denote $\mathrm{Spec}(\hat{\boldsymbol{\mathfrak{X}}}(\mathcal{M}, n))$ by $T^*[n]\mathcal{M}$ and have, tautologically,

$$\hat{\mathbf{C}}^\infty(T^*[n]\mathcal{M}) = \hat{\boldsymbol{\mathfrak{X}}}(\mathcal{M}, n), \tag{2.7}$$

and, noncanonically,

$$T^*[n]\mathcal{M} \simeq E \oplus T^*[n]M \oplus E^*[n]. \tag{2.8}$$

**Remark 2.13** (Multidifferential operators). Multidifferential operators may be defined as in 2.1.5. We will use the notations $\boldsymbol{\mathcal{D}}(\mathcal{M})$ and $\hat{\boldsymbol{\mathcal{D}}}(\mathcal{M})$ for the DGLAs $\mathcal{D}(\mathbf{C}^\infty(\mathcal{M}))$ and $\mathcal{D}(\hat{\mathbf{C}}^\infty(\mathcal{M}))$. The HKR maps $\boldsymbol{\mathfrak{X}}(\mathcal{M}) \to \boldsymbol{\mathcal{D}}(\mathcal{M})$ and $\hat{\boldsymbol{\mathfrak{X}}}(\mathcal{M}) \to \hat{\boldsymbol{\mathcal{D}}}(\mathcal{M})$ are quasiisomorphisms of differential complexes [17] (see also [18]).

**2.4.1. Poisson structures.** A smooth graded manifold $\mathcal{M}$ is called a graded Poisson manifold of degree $n$ if $\hat{\mathbf{C}}^\infty(\mathcal{M})$ is endowed with a bracket that makes it into an $n$-Poisson algebra. By (2.7), for every smooth graded manifold $\mathcal{M}$, $T^*[n]\mathcal{M}$ is a Poisson manifold of degree $n$ in a canonical way. As a Poisson bracket is a graded biderivation, an $n$-Poisson structure on $\hat{\mathbf{C}}^\infty(\mathcal{M})$ determines an element $\pi$ of $(S^2_{\hat{\mathbf{C}}^\infty(\mathcal{M})}(\mathrm{Der}(\hat{\mathbf{C}}^\infty(\mathcal{M}))[-1-n]))_{2+n}$. The Jacobi identity for the Poisson bracket is then equivalent to the equation $[\pi, \pi] = 0$. A bivector field of degree $-n$ satisfying this equation will be called an $n$-Poisson bivector field. The Poisson bracket of two functions $f$ and $g$ may then be recovered as the derived bracket

$$\{f, g\} = [[f, \pi], g], \tag{2.9}$$

where $f$ and $g$ are regarded on the r.h.s. as 0-vector fields.

[6] The equation holds also for an infinite-dimensional graded vector space $V$, if one works from the beginning in terms of coalgebras instead of algebras of functions so as to avoid taking double duals.

If the $n$-Poisson structure of a graded Poisson manifold is nondegenerate, we speak of a graded symplectic manifold of degree $n$.

So $T^*[n]M$ is a graded symplectic manifold of degree $n$ in a canonical way.[7] We call (anti)symplectomorphism between two graded symplectic manifolds a morphism of the underlying smooth graded manifolds that yields an (anti) isomorphism of the Poisson algebras of functions. We have the following fundamental

**Theorem 2.14** (Legendre mapping [34]). *Let $E$ be a graded vector bundle. Then $T^*[n]E$ is canonically antisymplectomorphic to $T^*[n](E^*[n])$ for all $n$.*

Observe that (2.8) implies that the two graded manifolds in the theorem are diffeomorphic. The additional statement is that there is a diffeomorphism preserving Poisson brackets up to a sign and that it is canonical (i.e., independent of the choice of connection used to prove (2.8)). For a proof, see [34].

**Remark 2.15.** The name "Legendre mapping" comes from the simplest instance [43] of this theorem in the category of manifolds, $T^*TM \simeq T^*T^*M$, which induces the usual Legendre transformation of functions. The generalization $T^*E \simeq T^*E^*$ is due to [32]. The explicit expression in coordinates of this map also suggests the name of "Fourier transformation" which is used in [17].

**2.5. Further readings.** In this short introduction we did not consider: local coordinates, the definition of graded manifolds as ringed spaces, differential and integral forms as well as a proper definition of graded submanifolds and of infinite-dimensional graded manifolds. We refer to [35] and references therein for further reading on graded manifolds. For supermanifolds, see also [4], [9], [21], [30], [44].

## 3. The BV formalism

We give here a presentation of the BV formalism [5], [23] (which is a generalization of the BRST [8], [42] formalism) based mainly on [38]. See also [2], [3], [13], [22], [24], [25].

**3.1. De Rham theory revisited.** Let $M$ be a smooth orientable manifold with a volume form $v$ and $\phi_v$ the isomorphism defined in Example 2.5. Define $\Delta_v := \phi_v^{-1} \circ \mathrm{d} \circ \phi_v$ where d is the exterior derivative. (Observe that $\Delta_v$ restricted to vector fields is just the divergence operator.) So $\Delta_v^2 = 0$. Since $\phi_v$ is not an algebra morphism, $\Delta_v$ is not a derivation; one can however show that

$$\Delta_v(XY) = \Delta_v(X)Y + (-1)^i X\Delta_v(Y) + (-1)^i[X, Y], \quad X \in \mathcal{X}^i(M),\ Y \in \mathcal{X}(M). \tag{3.1}$$

[7] It may be proved [38] that every graded symplectic manifold of degree $2k + 1$ is isomorphic to some $T^*[2k + 1]\mathcal{M}$ with canonical symplectic structure.

Since $\phi_v(X)$ is a differential form, it is natural to integrate it on a submanifold of the corresponding degree. Stokes' Theorem may then be reformulated by saying that the integral vanishes if $X$ is $\Delta_v$-exact, and that it is invariant under cobordisms if $X$ is $\Delta_v$-closed. Using the language of smooth graded manifolds as in Example 2.5, we then have the

**Theorem 3.1.** *Let $v$ be a volume form on $M$ and $X$ a function on $T^*[k]M$, $k$ odd. Then:*

1. $\int_{L_C} X\sqrt{\rho_v} = \int_{L_{C'}} X\sqrt{\rho_v}$ *for every two cobordant submanifolds $C$ and $C'$ of $M$ iff $X$ is $\Delta_v$-closed.*
2. $\int_{L_C} X\sqrt{\rho_v} = 0$ *for every $C$ iff $X$ is $\Delta_v$-exact.*

Let $Q_X := [X, \ ]$ denote the Hamiltonian vector field of $X \in \mathbf{C}^\infty(T^*[k]M) \simeq \mathfrak{X}(M, k)$, $k$ odd. Using (3.1) and Stokes' Theorem, one easily has the following characterization of $\Delta_v$ in terms of the canonical symplectic structure of $T^*[k]M$:

**Theorem 3.2.** $\Delta_v X = \frac{1}{2}\mathrm{div}_{\rho_v} Q_X$ *for every volume form $v$.*

By Lemma 2.6, we know that every Berezinian form on $T^*[k]M$ may be written, up to a constant, as $\mathrm{e}^\sigma \rho_v =: \rho_v^\sigma$ for some volume form $v$ and some function $\sigma$. We write $\sqrt{\rho_v^\sigma} := \mathrm{e}^{\frac{\sigma}{2}}\sqrt{\rho_v}$. By Theorem 3.1, $\int_{L_C}\sqrt{\rho_v^\sigma}$ is the same for all cobordant submanifolds iff $\mathrm{e}^{\frac{\sigma}{2}}$ is $\Delta_v$-closed. Assuming for simplicity $\sigma$ to be even, by Theorem 3.2 and Remark 2.8, one can show that this is the case iff

$$\Delta_v\sigma + \frac{1}{4}[\sigma, \sigma] = 0. \tag{3.2}$$

Given a solution $\sigma$ of this equation, one can define a new coboundary operator $\Omega_{v,\sigma} := \Delta_v + \frac{1}{2}Q_\sigma$. Remark that $\Omega X_{v,\sigma} = \mathrm{e}^{-\frac{\sigma}{2}}\Delta_v(\mathrm{e}^{\frac{\sigma}{2}}X)$. Thus, multiplication by $\mathrm{e}^{\frac{\sigma}{2}}$ is an invertible chain map $(\mathbf{C}^\infty(T^*[k]M), \Omega_{v,\sigma}) \to (\mathbf{C}^\infty(T^*[k]M), \Delta_v)$ and the two cohomologies are isomorphic. Moreover, Theorem 3.1 is still true if one replaces $(\rho_v, \sqrt{\rho_v}, \Delta_v)$ by $(\rho_v^\sigma, \sqrt{\rho_v^\sigma}, \Omega_{v,\sigma})$.

**3.2. The general BV formalism.** Even though the above setting is all we need in the present paper, for completeness we give an overview of the general results of [38]. For this one needs the notion of submanifold of a graded manifold as well as notions of symplectic geometry on graded manifolds which we are not going to introduce here.

**Theorem 3.3.** *Let $k$ be an odd integer. Then:*

1. *Theorem* 3.1 *holds if $M$ is a graded manifold and $v$ a Berezinian form.*
2. *Every graded symplectic manifold of degree $k$ is symplectomorphic to some $T^*[k]M$ with canonical symplectic form.*

3. *There is a canonical way* (*up to a sign*) *of restricting a Berezinian form* $\rho_v$ *on* $T^*[k]M$ *to a Berezinian form denoted by* $\sqrt{\rho_v}$ *on a Lagrangian submanifold.*

4. *Every Lagrangian submanifold* $L$ *of* $T^*[k]M$ *may be deformed to a Lagrangian submanifold of the form* $L_C$*, with* $C$ *a submanifold of* $M$.

5. *If* $X$ *is* $\Delta_v$*-closed, then* $\int_L X\sqrt{\rho_v} = \int_{L'} X\sqrt{\rho_v}$ *if* $L$ *may be deformed to* $L'$.

6. *If* $X$ *is* $\Delta_v$*-exact, then* $\int_L X\sqrt{\rho_v} = 0$ *for every Lagrangian submanifold* $L$.

**3.2.1. Generating functions.** To do explicit computations, it is useful to describe the Lagrangian submanifold in terms of generating functions. Generalizing concepts from symplectic geometry to graded manifolds, one sees that the graph of the differential of a function of degree $k$ on $M$ is a Lagrangian submanifold of $T^*[k]M$. Such a function is called a generating function. However, Lagrangian submanifolds of this form project onto $M$; so certainly a conormal bundle cannot be represented this way.

A slightly more general setting is the following. We assume here some knowledge of symplectic geometry (see e.g. [6]) and generalize a classical construction. Let $U$ be an auxiliary graded manifold, and let $f$ be a function of degree $k$ on $M \times U$. Let $\Sigma$ be the $U$-critical set of $f$; i.e., the subset $M \times U$ where the differential of $f$ along $U$ vanishes. Assume $\Sigma$ to be a submanifold and let $\phi\colon \Sigma \to T^*M$ be defined by $(x, u) \mapsto (x, \mathrm{d}f(x, u))$. Then $\phi$ is a Lagrangian immersion whose image we denote by $L(f)$.

For example, if $C$ is a submanifold of $M$ defined by global regular constraints $\phi_1, \ldots, \phi_r$, with $\phi_j$ of degree $n_j$, we may take $U := \bigoplus_{j=1}^r \mathbb{R}[n_j - k]$ and define $\Psi = \sum_j \beta^j \phi_j$, where $\beta_j$ is the coordinate on $\mathbb{R}[n_j - k]$. It turns then out that $L(\Psi) = N^*[k]C$.[8] We regard now $\Psi$ as a function on $\tilde{M} := M \times U \times U[-k]$ and denote by $L_\Psi$ the graph of its differential. On $U \times U[-k]$, we choose the Lebesgue measure for the even components and a pure Berezinian form for the odd ones. We denote by $\tilde{v}$ the Berezinian form on $\tilde{M}$ obtained by this times $\rho_v$. Finally, let $u$ be the pairing between $U$ and $U^*$ regarded as a function of degree zero on $U[-k] \times U^*[k]$ and hence, by pullback, on $T^*[k]\tilde{M}$. Then a simple computation (using the Fourier representation of the delta function) shows that

$$\int_{N^*[k]C} X\mathrm{e}^{\frac{\sigma}{2}}\sqrt{\rho_v} = \int_{L_\Psi} X\mathrm{e}^{\frac{\sigma}{2}+\mathrm{i}u}\sqrt{\rho_{\tilde{v}}}.$$

Observe that deforming $\Psi$ just deforms the Lagrangian submanifold (which in general will no longer be a conormal bundle) but leaves the result unchanged.

**3.3. BV notations.** The BV formalism consists of the above setting with $k = -1$ (for historical reasons). The $-1$-Poisson bracket is called BV bracket and usually denoted

[8] In the absence of global regular constraints, conormal bundles may be described by a further generalization of generating functions, the so-called Morse families. See, e.g., [6].

by $(\ ,\ )$. The coboundary operator $\Delta_v$ is called the BV Laplacian, has degree 1 and, as $v$ is fixed, is usually simply denoted by $\Delta$. A solution $\sigma$ to (3.2) is usually written as $\sigma = 2\frac{\mathrm{i}}{\hbar}\mathsf{S}$, where $\mathsf{S}$ is called the BV action and satisfies the so-called "quantum master equation" (QME) $(\mathsf{S},\mathsf{S}) - 2\mathrm{i}\hbar\Delta\mathsf{S} = 0$. Here $\hbar$ is a parameter and $\mathsf{S}$ is allowed to depend on $\hbar$. If $\mathsf{S}$ is of degree 0, as it is usually assumed, then $Q_{\mathsf{S}}$ is of degree 1. The coboundary operator $\Omega_{v,\sigma}$ is then also homogeneous of degree 1. Setting $\Omega := -\mathrm{i}\hbar\Omega_{v,\sigma}$, we have $\Omega = Q_{\mathsf{S}} - \mathrm{i}\hbar\Delta$. An $\Omega$-closed element $\mathcal{O}$ is called an observable, and its expectation value

$$\langle\mathcal{O}\rangle := \frac{\int_L \mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathsf{S}}\,\mathcal{O}\,\sqrt{\rho_v}}{\int_L \mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathsf{S}}\,\sqrt{\rho_v}} \tag{3.3}$$

is invariant under deformations of $L$. The choice of an $L$ goes under the name of gauge fixing.[9] Expectation values of $\Omega$-exact observables vanish, but they may lead to interesting relations called Ward identities.

**Remark 3.4.** One often assumes $\hbar$ to be "small." Actually, one even takes $\mathsf{S}$ to be a formal power series in $\hbar$, $\mathsf{S} = \sum_{i=0}^{\infty}\hbar^i\mathsf{S}_i$. Then $\mathsf{S}_0$ satisfies the "classical master equation" (CME) $(\mathsf{S},\mathsf{S}) = 0$ and $Q_{\mathsf{S}_0}$ is a coboundary operator (sometimes called the BRST operator). One may look for solutions of the QME starting from a solution $\mathsf{S}_0$ of the CME. One easily sees that there is a potential obstruction to doing this (the so-called anomaly) in the second cohomology group of $Q_{\mathsf{S}_0}$.

**Remark 3.5.** An observable $\mathcal{O}$ of degree zero may also be thought of as an infinitesimal deformation of the BV action, for $\mathsf{S} + \varepsilon\mathcal{O}$ then satisfies the CME up to $\varepsilon^2$. For this to be a finite deformation, we should also assume $(\mathcal{O},\mathcal{O}) = 0$.

**3.4. Applications.** Suppose that the integral of $\mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathsf{S}}$ along a Lagrangian submanifold $L$ is not defined, but that it is enough to deform $L$ a little bit for the integral to exist. Then one defines the integral along $L$ as the integral along a deformed Lagrangian submanifold $L'$. For a given cobordism class of deformations, the integral does not depend on the specific choice of $L'$ if $S$ is assumed to satisfy the QME. This is really analogous to the definition of the principal part of an integral [22].

The typical situation is the following: One starts with a function $S$ defined on some manifold $\mathcal{M}$. One assumes there is a (nonnecessarily integrable) distribution on $\mathcal{M}$ – the "symmetries" – under which $S$ is invariant. One then adds odd variables of degree 1 (the generators of the distribution, a.k.a. the ghosts) defining a graded manifold $\tilde{\mathcal{M}}$ which fibers over $\mathcal{M}$ and is endowed with a vector field $\delta$ that describes the distribution. Then one tries to extend $S$ to a solution $\mathsf{S}_0 \in \mathbf{C}^\infty(T^*[-1]\tilde{\mathcal{M}})$ of the CME such that $Q_{\mathsf{S}_0}$ and $\delta$ are related vector fields. Under the assumption that the original distribution is integrable on the subset (usually assumed to be a submanifold)

[9] This is usually done as explained in 3.2.1 by using an auxiliary space and a generating function $\Psi$ which is in this case of degree $-1$ and is called the gauge-fixing fermion.

of critical points of $S$, one can show that this is possible under some mild regularity assumptions [5]. The next step is to find a solution of the QME as in Remark 3.4 if there is no anomaly.

Because of the invariance of $S$, the integral of $\mathrm{e}^{\frac{\mathrm{i}}{\hbar}S}$ on $\mathcal{M}$ will diverge (if the symmetry directions are not compact). On the other hand, if we integrate over $\tilde{\mathcal{M}}$ we also have zeros corresponding to the odd directions which we have introduced and along which $S$ is constant. If we introduce all generators, we have as many zeros as infinities, so there is some hope to make this ill-defined integral finite. This is actually what happens if we find a solution of the QME as in the previous paragraph and integrate on a different Lagrangian submanifold of $T^*[-1]\tilde{\mathcal{M}}$ than its zero section $\tilde{\mathcal{M}}$.

Given a function $\mathcal{O}$ on $\mathcal{M}$, it makes sense to define its expectation value as in (3.3) if there is an observable $\boldsymbol{\mathcal{O}}$ whose restriction to $\mathcal{M}$ is $\mathcal{O}$.

**Remark 3.6** (Field theory). In field theory one considers integrals of the form (1.8) with $\mathcal{M}$ infinite dimensional. Integration around critical points is defined by expanding the non quadratic part of $S$ and evaluating Gaussian expectation values. If there are symmetries, the critical points are degenerate and one cannot invert the quadratic form. One then operates as above getting an integral with the quadratic part of the BV action nondegenerate, so one can start the perturbative expansion.[10] This is not the end of the story since two problems arise. The first is that the formal evaluation of the Gaussian expectation values leads to multiplying distributions. The consistent procedure for overcoming this problem, when possible, goes under the name of renormalization. The second problem is that, in the absence of a true measure, there is no divergence operator and thus no well-defined BV Laplacian $\Delta$. This is overcome by defining $\Delta$ appropriately in perturbation theory. On the other hand, the BV bracket is well-defined (on a large enough class of functions). In the present paper the field theory is so simple that renormalization is (almost) not needed, so we will not talk about it. On the other hand, it makes sense [14] to assume that $\Delta$ exists and vanishes on the local functionals we are going to consider, while on products thereof one uses (3.1).

## 4. BV 2D TFT

We go back now to our original problem described in the Introduction. This may also be regarded as a continuation of our presentation in [10, Part III].

**4.1. The BV action.** We start by considering the TFT with action (1.1) and symmetries (1.2). We promote the generators $\beta$ of the symmetries to odd variables of degree 1; i.e., we define $\tilde{\mathcal{M}}_1 = \mathcal{M}_1 \oplus \Omega^0(\Sigma)[1]$ and the vector field $\delta$ by its action on the linear function $\xi$, $\eta$ and $\beta$: $\delta\xi = 0$, $\delta\eta = \mathrm{d}\beta$, $\delta\beta = 0$. Using integration on $\Sigma$, we identify $T^*[-1]\tilde{\mathcal{M}}_1$ with $\tilde{\mathcal{M}}_1 \oplus \Omega^2(\Sigma)[-1] \oplus \Omega^1(\Sigma)[-1] \oplus \Omega^2(\Sigma)[-2]$ and denote

[10] In order to have Gaussian integration on a vector space, one defines integration along the chosen Lagrangian submanifold via a generating function as explained in 3.2.1 and in footnote 9.

the new coordinates, in the order, by $\xi^+$, $\eta^+$ and $\beta^+$. We introduce the "superfields" $\boldsymbol{\xi} = \xi + \eta^+ + \beta^+$, $\boldsymbol{\eta} = \beta + \eta + \xi^+$, and define

$$\mathsf{S}(\boldsymbol{\xi}, \boldsymbol{\eta}) := \int_\Sigma \boldsymbol{\eta}\, \mathrm{d}\boldsymbol{\xi}, \tag{4.1}$$

where by definition the integration selects the 2-form. It is not difficult to see that $\mathsf{S}$ satisfies the master equation and $\mathsf{S}_{|\mathcal{M}_1} = S$. Moreover, the action of $Q_\mathsf{S}$ on the coordinate functions may be summarized in

$$Q_\mathsf{S}\boldsymbol{\xi} = \mathrm{d}\boldsymbol{\xi}, \quad Q_\mathsf{S}\boldsymbol{\eta} = \mathrm{d}\boldsymbol{\eta}. \tag{4.2}$$

So $Q_\mathsf{S}$ and $\delta$ are related vector fields.

By (2.3), we may regard $\boldsymbol{\xi}$ as an element of $\Omega(\Sigma)$ and $\boldsymbol{\eta}$ as an element of $\Omega(\Sigma)[1]$. As $\Omega(\Sigma) = \mathbf{C}^\infty(T[1]\Sigma)$, by Remark 2.12 at the end we may further identify $\Omega(\Sigma)$ with $\mathrm{Map}(T[1]\Sigma, \mathbb{R})$ and $\Omega(\Sigma)[1]$ with $\mathrm{Map}(T[1]\Sigma, \mathbb{R}[1])$ or, equivalently, with $\mathrm{Map}(T[1]\Sigma, \mathbb{R}^*[1])$. The latter choice is more appropriate in view of (4.1) where we pair $\boldsymbol{\xi}$ with $\boldsymbol{\eta}$. By Remark 2.12 at the end again, we have eventually the identification $T^*[-1]\tilde{\mathcal{M}}_1 \simeq \mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R})$, where we have identified $\mathbb{R} \oplus \mathbb{R}^*[1]$ with $T^*[1]\mathbb{R}$ (by the results of Example 2.11 with $E = \mathbb{R}$ as a vector bundle over a point). This is actually the viewpoint taken in [1] (see also [15]). Finally, observe that we may also regard $T^*[-1]\tilde{\mathcal{M}}_1$ as $\mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R}[0])$ (which is actually a submanifold) if we wish to consider formal power series in the coordinate functions.

The ill-defined integration on $\tilde{\mathcal{M}}_1$ is now replaced by a well-defined (in the sense of perturbation theory) integration over another Lagrangian submanifold $L$ of $T^*[-1]\tilde{\mathcal{M}}[-1]$. For example, as in [14], we may take $L = N^*[-1]C$ where $C$ is the submanifold of $\tilde{\mathcal{M}}$ defined as the zero locus of $\mathrm{d} * \eta$, where the Hodge-star operator is defined upon choosing a volume form on $\Sigma$.

**4.2. The superpropagator.** The main object appearing in the explicit evaluation of expectation values of functions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ is the "superpropagator" $\langle \boldsymbol{\xi}(z)\boldsymbol{\eta}(w)\rangle$, where $z$ and $w$ are points in $\Sigma$. Independently of the choice of gauge fixing, we have the Ward identity

$$\begin{aligned} 0 &= \langle \Omega(\boldsymbol{\xi}(z)\boldsymbol{\eta}(w))\rangle = \langle Q_\mathsf{S}(\boldsymbol{\xi}(z)\boldsymbol{\eta}(w))\rangle - \mathrm{i}\hbar\langle \Delta(\boldsymbol{\xi}(z)\boldsymbol{\eta}(w))\rangle \\ &= \mathrm{d}\langle(\boldsymbol{\xi}(z))\boldsymbol{\eta}(w)\rangle - \mathrm{i}\hbar\langle(\boldsymbol{\xi}(z), \boldsymbol{\eta}(w))\rangle = \mathrm{d}\langle(\boldsymbol{\xi}(z))\boldsymbol{\eta}(w)\rangle - \mathrm{i}\hbar\delta(z, w), \end{aligned}$$

where we assumed $\Delta(\boldsymbol{\xi}(z)) = \Delta(\boldsymbol{\eta}(w)) = 0$ (which is consistent with perturbation theory) and $\delta$ denotes the delta distribution (regarded here as a distributional 2-form). Thus, we get the fundamental identity[11]

$$\mathrm{d}\langle(\boldsymbol{\xi}(z)\boldsymbol{\eta}(w)\rangle = \mathrm{i}\hbar\delta(z, w). \tag{4.3}$$

[11] This method for deriving properties of the superpropagator just in terms of Ward identities works also for the higher-dimensional generalization of this TFT [19].

The restriction of the superpropagator to the configuration space $C_2(\Sigma) := \{(z, w) \in \Sigma \times \Sigma : z \neq w\}$ is then a closed, smooth 1-form. Namely, we set $i\hbar\theta(z, w) := \langle \boldsymbol{\xi}(z)\boldsymbol{\eta}(w)\rangle$, $(z, w) \in C_2(\Sigma)$. Then $\theta \in \Omega^1(C_2(\Sigma))$ and $d\theta = 0$. We call it the propagator 1-form. The delta distribution in (4.3) implies that $\int_\gamma \theta = 1$ where $\gamma$ is generator of the singular homology of $C_2(\Sigma)$ (viz., $\gamma$ is a loop of $w$ around $z$ or vice versa). Observe that $\theta$ is defined up to an exact 1-form. Different choices of gauge fixing just correspond to different, but cohomologous, choices of $\theta$.

If $\partial\Sigma \neq \emptyset$, we have to choose boundary conditions. Repeating the considerations in the Introduction, we see that there are two possible boundary conditions compatible with (4.2); viz.:

$$\begin{aligned} &\text{Neumann boundary conditions:} && \iota^*_{\partial\Sigma}\boldsymbol{\eta} = 0, && \text{(N)}\\ &\text{Dirichlet boundary conditions:} && \iota^*_{\partial\Sigma}\boldsymbol{\xi} = 0, && \text{(D)} \end{aligned}$$

For $\partial\Sigma = \emptyset$, the BV action (4.1) is invariant under the exchange of $\boldsymbol{\eta}$ with $\boldsymbol{\xi}$. This implies that $\psi^*\theta = \theta$ with $\psi(z, w) = (w, z)$.[12] For $\partial\Sigma \neq \emptyset$, we denote by $\theta_N$ and $\theta_D$ the propagator 1-forms corresponding to N and D boundary conditions, respectively. These 1-forms have to satisfy in addition boundary conditions. Let $\partial_i C_2(\Sigma) = \{(z_1, z_2) \in C_2(\Sigma) : z_i \in \partial\Sigma\}$ and $\iota_i$ the inclusion of $\partial_i C_2(\Sigma)$ into $C_2(\Sigma)$. Then we have $\iota_1^*\theta_D = 0$ and $\iota_2^*\theta_N = 0$. These 1-forms are no longer invariant under the involution $\psi$ defined above; they are instead related by it: viz., $\psi^*\theta_N = \theta_D$.

**4.3. Duality.** Exchanging the superfields has a deeper meaning. Observe that the 0-form component $\xi$ of $\boldsymbol{\xi}$ is an ordinary function (of degree zero), while the 0-component form $\beta$ of $\boldsymbol{\eta}$ has been assigned degree 1 and has values in $\mathbb{R}^*$. So, when we make this exchange, we are actually exchanging, loosely speaking, a map $\xi\colon \Sigma \to \mathbb{R}[0]$ with a map $\beta\colon \Sigma \to \mathbb{R}^*[1]$. In exchanging the superfields, we are then actually performing the canonical symplectomorphism $\mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R}[0]) \to \mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R}^*[1])$ which is induced by the canonical symplectomorphism $T^*[1]\mathbb{R}[0] \to T^*[1]\mathbb{R}[1]$, a special case of the Legendre mapping of Theorem 2.14. If we now take the graded vector space $\mathbb{R}[k]$ as target, the superfield exchange is a symplectomorphism $\mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R}[k]) \to \mathrm{Map}(T[1]\Sigma, T^*[1]\mathbb{R}^*[1-k])$. In conclusion, the TFT with target $\mathbb{R}[k]$ is equivalent to the TFT with target $\mathbb{R}^*[1-k]$ is $\Sigma$ has no boundary, whereas, if $\Sigma$ has a boundary, the TFT with target $\mathbb{R}[k]$ and N boundary conditions is equivalent to the TFT with target $\mathbb{R}^*[1-k]$ and D boundary conditions. Thus, upon choosing the target appropriately, one may always assume to have only N boundary conditions.

**4.4. Higher-dimensional targets.** We may allow a higher-dimensional target as in (1.3) or in (1.5), and it makes sense for it to be a graded vector space or a graded manifold $M$. Now the space of fields may be identified with $\mathrm{Map}(T[1]\Sigma, T^*[1]M)$. For

[12] The cohomology class of a propagator 1-form is necessarily $\psi$-invariant. The stronger condition is that it is $\psi$-invariant without passing to cohomology.

simplicity, assume the target to be a graded vector space $V$ (which is anyway the local version of the general case). Upon choosing a graded basis $\{e_I\}$ and its dual basis $\{e^I\}$, we may consider the components $\boldsymbol{\xi}^i$ and $\boldsymbol{\eta}_i$ of the superfields. The superpropagator may then be computed as $\langle \boldsymbol{\xi}^I(z)\boldsymbol{\eta}_J(w)\rangle = \mathrm{i}\hbar\theta(z,w)\delta^I_J$, $(z,w) \in C_2(\Sigma)$, where $\theta$ is the 1-form propagator of the TFT with target $\mathbb{R}$. Again we are allowed to exchange superfields, but we may decide to exchange only some of them. Let $V = W_1 \oplus W_2$. A superfield exchange corresponding to $W_2$-components establishes a symplectomorphism $\mathrm{Map}(T[1]\Sigma, T^*[1](W_1 \oplus W_2) \simeq \mathrm{Map}(T[1]\Sigma, T^*[1](W_1 \oplus W_2^*[1])$. If we have N boundary conditions on the $W_1$-components and D boundary conditions on the $W_2$-components, the exchange yields a theory with only N boundary conditions.

If we work with target a graded manifold $M$ and D boundary conditions on a graded submanifold $C$, the perturbative expansion actually sees as target the graded submanifold $N[0]C$ of $M$ (as a formal neighborhood of $C$). As a consequence of the previous considerations, this is the same as the TFT with target $N^*[1]C$ and N boundary conditions. This case has been studied in [16], [17].

**4.4.1. Assumptions.** From now on we assume that $\Sigma$ is the disk and that on its boundary $S^1$ we put N boundary conditions. We also choose a point $\infty \in S^1$ and fix the map $X$ to take the value $x \in M$ at $\infty$. By setting $X = x + \xi$ we identify the theory with target $M$ with the theory with target the graded vector space $T_x[0]M$. The superfield $\boldsymbol{\xi} \in \mathrm{Map}(T[1]\Sigma, T_x[0]M)$ is then assumed to vanish at $\infty$.

**4.5. Ward identities and formality theorem.** There exists a class of interesting observables associated to multivector fields on the target. For simplicity we assume the target to be a graded vector space $V$, make the identification (2.2) and use a graded basis. So, for a $k$-vector field $F \in \mathfrak{X}(V)$, we define

$$\mathsf{S}_F(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{1}{k!}\int_\Sigma F^{i_1\dots i_k}(\boldsymbol{\xi})\boldsymbol{\eta}_{i_1}\dots\boldsymbol{\eta}_{i_k}. \tag{4.4}$$

Since $Q_{\mathsf{S}}\mathsf{S}_F = \frac{1}{k!}\int_{\partial\Sigma} F^{i_1\dots i_k}(\boldsymbol{\xi})\boldsymbol{\eta}_{i_1}\dots\boldsymbol{\eta}_{i_k}$., we have defined an observable unless $F$ is a 0-vector field (i.e., a function), for one may show [14] that it is consistent to assume $\Delta\mathsf{S}_F = 0$. We will call observables of this kind bulk observables. By linear extension, we may associate a bulk observable to every element $F \in \hat{\mathfrak{X}}(V)$. If $F$ is of total degree $f$, then $\mathsf{S}_F$ is of degree $f-2$. One may also show [14] (see also [15]) that $(\mathsf{S}_F, \mathsf{S}_G) = \mathsf{S}_{[F,G]}$ for any two multivector fields $F$ and $G$. Another interesting class of observables is associated to functions on the target. Given a function $f$ and a point $u \in \partial\Sigma$, we set $\boldsymbol{\mathcal{O}}_{f,u}(\boldsymbol{\xi}, \boldsymbol{\eta}) = f(\boldsymbol{\xi}(u)) = f(\xi(u))$. Since $Q_{\mathsf{S}}\boldsymbol{\mathcal{O}}_{f,u} = 0$ as $u$ is on the boundary, since the difference $\boldsymbol{\mathcal{O}}_{f,u} - \boldsymbol{\mathcal{O}}_{f,u'}$ is equal $Q_{\mathsf{S}}\int_u^{'u} f(\boldsymbol{\xi})$ and since one may consistently set to zero $\Delta$ applied to functions of $\boldsymbol{\xi}$ only, we have defined new observables, which we will call boundary observables, in which the choice of $u$ is immaterial.

A product of observables is in general not an observable (since $\Omega$ is not a derivation). A product which is however an observable is $\boldsymbol{\mathcal{O}}(F; f_1, \dots, f_k)_{u_1,\dots u_k} :=$

$\mathsf{S}_F\boldsymbol{\mathcal{O}}_{f_1,u_1}\dots\boldsymbol{\mathcal{O}}_{f_k,u_k}$, where $F$ is a $k$-vector field, $k>0$, the $f_i$s are functions and the $u_i$s are ordered points on the boundary. The expectation value may easily be computed [14] and one gets $\langle\boldsymbol{\mathcal{O}}(F;f_1,\dots,f_k)_{u_1,\dots u_k}\rangle = \mathrm{HKR}(F)(f_1\otimes\cdots\otimes f_k)$. More generally, one may define

$$\boldsymbol{\mathcal{O}}(F_1,\dots,F_m;f_1,\dots,f_k)_{u_1,\dots u_k} := \mathsf{S}_{F_1}\dots\mathsf{S}_{F_m}\boldsymbol{\mathcal{O}}_{f_1,u_1}\dots\boldsymbol{\mathcal{O}}_{f_k,u_k}.$$

One may show [14] that the expectation value of $\boldsymbol{\mathcal{O}}(F_1,\dots,F_m;f_1,\dots,f_k)_{u_1,\dots u_k}$ may be regarded as a multidifferential operator $U_m(F_1,\dots,F_m)$ acting on $f_1\otimes\cdots\otimes f_k$. This way one defines multilinear maps $U_m$s from $\boldsymbol{\mathcal{X}}$ to $\boldsymbol{\mathcal{D}}$. However, the explicit form of the multidifferential operators will depend on the chosen gauge fixing as $\boldsymbol{\mathcal{O}}(F_1,\dots,F_m;f_1,\dots,f_k)_{u_1,\dots u_k}$ is not an observable in general. One may get very interesting identities relating the $U_m$s by considering the Ward identities

$$0=\langle\Omega\boldsymbol{\mathcal{O}}(F_1,\dots,F_m;f_1,\dots,f_k)_{u_1,\dots u_k}\rangle. \tag{4.5}$$

One may show [14] that the various contribution of the r.h.s. correspond to collapsing in all possible ways some of the bulk observables together with some of the boundary observables (with consecutives $u$s). As a result one gets relations among the $U_m$s. To interpret them, we have to introduce some further concepts.

**Definition 4.1.** An $L_\infty$-algebra[13] [29], [41] is a graded vector space $V$ endowed with operations (called multibrackets) $L_k\in\mathrm{Hom}_1(S^kV,V)$, $k\in\mathbb{N}$, satisfying for all $n\geq 0$ and for all $v_1,\dots,v_n\in V$

$$\sum_{k+l=n}\ \sum_{\sigma\in(k,l)\text{-shuffles}}\mathrm{sign}(\sigma)L_{l+1}(L_k(v_{\sigma(1)},\dots,v_{\sigma(k)}),v_{\sigma(k+1)},\dots,v_{\sigma(n)})=0,$$

where a $(k,l)$-shuffle is a permutation on $k+l$ elements such that $\sigma(1)<\cdots<\sigma(k)$ and $\sigma(k+1)<\cdots<\sigma(k+l)$, while the sign of the permutation $\sigma$ is defined by $v_{\sigma(1)}\dots v_{\sigma(n)}=\mathrm{sign}(\sigma)v_1\dots v_n$ in $S^kV$. We call flat an $L_\infty$-algebra with $L_0=0$.

In a flat $L_\infty$-algebra, $L_1$ is a coboundary operator. We denote by $H(V)$ the $L_1$-cohomology. Observe that $H(V)[-1]$ acquires a DGLA structure.

For $V$ finite dimensional, we may identify $\mathrm{Hom}_1(SV,V)$ with $(SV^*\otimes V)_1$ and so with $\mathfrak{X}(V)_1$. An $L_\infty$-algebra on $V$ is then the same as the data of a "cohomological vector field" (i.e., a vector field of degree 1 that squares to zero). The same holds in the infinite-dimensional case if one defines things appropriately.

**Example 4.2.** A (D)GLA $\mathfrak{g}$ may be regarded as a flat $L_\infty$-algebra by setting $V=\mathfrak{g}[1]$ and defining $L_k$ to be the Lie bracket for $k=2$ (and the differential for $k=1$), while all other $L_k$s are set to zero.

One may introduce the category of $L_\infty$-algebras by defining an $L_\infty$-morphism from $V$ to $W$ to be a sequence of morphisms $SV\to W$ with appropriate relations between the two sets of multibrackets. We do not spell these relations here. They essentially state that there is a morphism $V\to W$ as (possibly infinite-dimensional) graded

[13] We follow here the sign conventions of [45].

manifolds such that the corresponding homological vector fields are related. We write $U\colon V \rightsquigarrow W$ for an $L_\infty$-morphisms with components $U_m \in \mathrm{Hom}_0(S^m V, W)$. An important properties of the definition is the following: If $V$ and $W$ are flat and $U_0 = 0$, then $U_1$ is a chain map. If $U_1$ induces an isomorphism in cohomology, one says that $U$ is an $L_\infty$-quasiisomorphism. If in addition $V$ has zero differential, $V[-1]$ is isomorphic as a GLA to $H(W)[-1]$, and one says that $W[-1]$ is formal. Finally we may interpret the Ward identities (4.5) in terms of the DGLAs $\hat{\boldsymbol{\mathcal{V}}}(M) := \hat{\boldsymbol{\mathcal{X}}}(M)[1]$ and $\hat{\boldsymbol{\mathcal{D}}}(M)$ as flat $L_\infty$-algebras:

**Theorem 4.3** (Formality Theorem). *There is an $L_\infty$-morphism $U\colon \hat{\boldsymbol{\mathcal{V}}}(M) \rightsquigarrow \hat{\boldsymbol{\mathcal{D}}}(M)$, with $U_1$ the HKR map. So $U$ is an $L_\infty$-quasiisomorphism and the DGLA $\hat{\boldsymbol{\mathcal{D}}}(M)$ is formal.*

The Ward identities are not a full proof of the theorem as all arguments using infinite-dimensional integrals have to be taken with care (e.g., we have always assumed that we can work with the BV Laplacian $\Delta$ which is actually not properly defined). They however strongly suggest that such a statement is true. One may check that this is the case by inspecting the finite-dimensional integrals (associated to the Feynman diagrams) appearing in the perturbative expansion. For $M$ an ordinary smooth manifold, the Formality Theorem has been proved by Kontsevich in [28]. For a proof when $M$ is a smooth graded manifold, see [17].

**4.6. Deforming the action: The Poisson sigma model.** As we observed in Remark 3.5, an observable of degree zero that commutes with itself may be used to deform the BV action. By considering bulk observables (4.4), we get a deformed BV action $\mathsf{S}_F^{\mathrm{def}} = \mathsf{S} + \varepsilon \mathsf{S}_F$ for every self-commuting $F = \sum_i F_i \in \boldsymbol{\mathcal{X}}(M)_2$, $F_i$ is an $i$-vector field, which does not contain a 0-vector field (i.e., $F_0 = 0$).

An element $x$ of degree one of a DGLA is called an MC (for Maurer–Cartan) element if $\mathrm{d}x + \frac{1}{2}[x, x] = 0$. So $F$ must be in particular an MC element in $\boldsymbol{\mathcal{V}}(M)$. A multivector field $F$ is completely characterized by its derived brackets

$$\begin{aligned}\lambda_i(a_1, \dots, a_i) &:= \mathrm{pr}\,([\,[\cdots[\,[F, a_1], a_2], \dots], a_i]) \\ &= [\,[\cdots[\,[F_i, a_1], a_2], \dots], a_i], \quad a_1, \dots, a_i \in \hat{\mathbf{C}}^\infty(M),\end{aligned}$$

where pr is the projection from $\hat{\boldsymbol{\mathcal{V}}}(M)$ onto the abelian Lie subalgebra $\hat{\mathbf{C}}^\infty(M)$. A consequence of the more general results in [45] is that $F$ is MC iff $(\hat{\mathbf{C}}^\infty(M), \lambda)$ is an $L_\infty$-algebra. The condition $F_0 = 0$ is precisely the condition that this $L_\infty$-algebra is flat. By construction the multibrackets $\lambda$ are multiderivations, so we call this $L_\infty$-algebra a $P_\infty$-algebra ($P$ for Poisson).

A particular case is when $F$ is a Poisson bivector field of degree zero. This is the only possibility if the target is an ordinary manifold. The only derived bracket is the Poisson bracket (2.9), and $\mathsf{S}_F^{\mathrm{def}}$ is the BV action of the so-called Poisson sigma model [27], [37]. Another particular case is when we start with an ordinary Poisson manifold $(P, \pi)$ and consider the Poisson sigma model with D boundary conditions

on a submanifold $C$. As discussed at the end of 4.4, this is the same as working with target $N^*[1]C$ and N boundary conditions. The Poisson bivector field $\pi$ induces, noncanonically, a Poisson bivector field $\tilde{\pi}$ on $N[0]C$ which in turns by the Legendre transform yields an MC element $F$ in $\hat{\boldsymbol{\mathcal{V}}}(N^*[1]C)$. As pointed out above, we need $F_0 = 0$. This is the case iff $C$ is a coisotropic submanifold [16]. A submanifold $C$ of a Poisson manifold $P$ is called coisotropic if its vanishing ideal $I$ is a Lie subalgebra of $(C^\infty(P), \{\ ,\ \})$.[14] The derived brackets on $\hat{\mathbf{C}}^\infty(N^*[1]C)$ yield the $L_\infty$-algebra studied in [33]. The zeroth $F_1$-cohomology group is the Poisson algebra $C^\infty(C)^I$ of $\{I,\ \}$-invariant functions on $C$. Hamiltonian vector fields of functions in $I$ define an integrable distribution on $C$. The leaf space $\underline{C}$ is called the reduction of $C$. If it is a manifold, $C^\infty(\underline{C}) = C^\infty(C)^I$.[15]

The expectation value of boundary observables in the deformed theory $\mathsf{S}_F^{\mathrm{def}}$ may easily be computed in perturbation theory by expanding $\exp(\varepsilon \mathsf{S}_F)$. As a result one has just to apply to the functions placed on the boundary the formal power series of multidifferential operator $U(\varepsilon F) := \sum_{k=1}^\infty \frac{\varepsilon^k}{k!} U_k(F, \dots, F)$.

If $\mathfrak{g}$ is a DGLA, by linearity one may extend the differential and the bracket to formal power series and so give $\varepsilon\mathfrak{g}[[\varepsilon]]$ the structure of a DGLA. If $x$ is an MC element in a GLA $\mathfrak{g}$, then $\varepsilon x$ is an MC element in $\varepsilon\mathfrak{g}[[\varepsilon]]$. An $L_\infty$-morphism $U\colon \mathfrak{g} \rightsquigarrow \mathfrak{h}$ between DGLAs $\mathfrak{g}$ and $\mathfrak{h}$ may be extended by linearity to formal power series as well. If $X$ is an MC element in $\varepsilon\mathfrak{g}[[\varepsilon]]$, then $U(X)$ is well defined in $\varepsilon\mathfrak{h}[[\varepsilon]]$ and it may be proved to be an MC element.

So $U(\varepsilon F)$ is an MC element in $\varepsilon\hat{\boldsymbol{\mathcal{D}}}(M)[[\varepsilon]]$. As shown in [17] such an MC element induces an $A_\infty$-structure on $\hat{\mathbf{C}}^\infty(M)[[\varepsilon]]$. This is the data of multibrackets $A_i$ (with $i$ arguments) satisfying relations analogous to those of an $L_\infty$-algebra but without symmetry requirements [40], [41]. If $A_0 = 0$, the $A_\infty$-algebra is called flat, $A_1$ is a differential for $A_2$, and the $A_1$-cohomology has the structure of an associative algebra. However, $A_0 = 0$ is not implied by $F_0 = 0$. In [17] it is proved that a potential obstruction to making the $A_\infty$-structure flat is contained in the second $F_1$-cohomology group. We call this potential obstruction the anomaly.

## 5. Applications

When the target $M$ is an ordinary manifold and $F$ is a Poisson bivector field, $C^\infty(M)$ is concentrated in degree zero, so the $A_\infty$-structure consists just of the bidifferential operator and is a genuine associative algebra structure. This is the original result by Kontsevich [28] that every Poisson bivector field defines a deformation quantization [7] of the algebra of functions.

[14] According to Dirac's terminology, $C$ is determined (locally) by first-class constraints.

[15] We discuss here deformations of the TFT $\mathsf{S}$, i.e, the Poisson sigma model with zero Poisson structure. If one drops the condition that the Poisson sigma model with D boundary conditions must be such a deformation, much more general submanifolds $C$ are allowed [11], [12].

A general method for studying certain submanifolds of so-called weak Poisson manifolds and their quantization has been suggested in [31]: one concocts a smooth graded manifold $M$ endowed with an MC element $F$, with $F_0 = 0$, to describe the problem, and then applies the $L_\infty$-quasiisomorphism $U$.

A particular case is the graded manifold $N^*[1]C$ associated to a coisotropic submanifold $C$, as described above. In the absence of anomaly, the method yields a deformation quantization of a Poisson subalgebra of $C^\infty(C)^I$ (or of the whole algebra if the first $F_1$-cohomology vanishes) [16], [17].

A second interesting case is that of a Poisson submanifold $P'$ of a Poisson manifold $P$. The inclusion map $\iota$ is then a Poisson map (i.e., $\iota^*$ is a morphism of Poisson algebras). One may then try to get deformation quantizations of $P$ and $P'$ together with a morphism of associative algebras that deforms $\iota^*$. The simpler case is when $P'$ is determined by regular constraints $\phi^1, \dots, \phi^k$. The Koszul resolution of $C^\infty(P')$ is obtained by introducing variables $\mu^1, \dots, \mu^k$ of degree $-1$ and defining a differential $\delta\mu^i = \phi^i$. We may interpret this differential as a cohomological vector field $Q$ on the graded manifold $M := P \times \mathbb{R}^k[-1]$. The Poisson bivector field $\pi$ on $P$ may also be regarded as a Poisson bivector field on $M$. We may put the two together defining $F = Q + \pi$, which is an MC element iff $[\pi, Q] = 0$, i.e., iff the $\phi^i$s are central. In this case $U(\varepsilon F)$ produces an $A_\infty$-algebra structure on $\mathbf{C}^\infty(M)[[\varepsilon]]$, which is flat since $\mathbf{C}^\infty(M)$ is concentrated in nonpositive degrees. Moreover, $\mathbf{C}^\infty(M)_0[[\varepsilon]] = C^\infty(P)[[\varepsilon]]$ inherits an algebra structure which turns out to give a deformation quantization of $P$. One may also verify that the zeroth $A_1$-cohomology group $H^0$ is a deformation quantization of $P'$ and that the projection $\mathbf{C}^\infty(M)_0[[\varepsilon]] \to H^0$, which is by construction an algebra morphism, is a deformation of $\iota^*$. By inspection of the explicit formulae, one may easily see that this construction is the same as the one proposed in [20], thus proving their conjecture. The more general case when the regular constraints $\phi^i$ are not central, may in principle be treated following [29] which shows the existence an MC element of the form $F = Q + \pi + O(\mu)$. Repeating the above reasoning does not solve the problem since in general the algebra $\mathbf{C}^\infty(M)_0[[\varepsilon]]$ is not associative. For this to be the case, one has to find corrections to $F$ such that in each term the polynomial degree in the $\mu^i$s is less or equal than the polynomial degree in the $\partial/\partial\mu^i$s.

A third interesting case is that of a Poisson map $J$ from a Poisson manifold $P$ to the dual of a Lie algebra $\mathfrak{g}$. Under certain regularity assumptions, $J^{-1}(0)$ is a coisotropic submanifold and may be quantized as described above. In practice, the formulae are not very explicit, even if $P$ is a domain in $\mathbb{R}^n$, for one has to choose adapted coordinates. A different approach is the following: First endow $P \times \mathfrak{g}^*$ with the unique Poisson structure which makes the projection $\mathrm{p}_1$ to $P$ Poisson, the projection $\mathrm{p}_2$ to $\mathfrak{g}^*$ anti-Poisson and such that $\{\mathrm{p}_2^* X, \mathrm{p}_1^* f\}_{P\times\mathfrak{g}^*} = p_1^*\{J_X, f\}_P$, for all $f \in C^\infty(P)$ and for all $X \in \mathfrak{g}$. The graph $G$ of $J$ is then a Poisson submanifold of $P \times \mathfrak{g}^*$, while $P \times \{0\}$ is coisotropic. Their intersection, diffeomorphic to $J^{-1}(0)$, turns out to be coisotropic in $G$. One then describes $G$ as the zero set of the regular constraints $\phi\colon P \times \mathfrak{g}^* \to \mathfrak{g}^*$, $(x, \alpha) \mapsto J(x) - \alpha$. Thus, applying the above construction,

one describes $G$ by an appropriate MC element $F$ on $M := P \times \mathfrak{g}^* \times \mathfrak{g}^*[-1]$ and realizes the quantization of $\underline{J^{-1}(0)}$ by the TFT with BV action $\mathsf{S}_F^{\mathrm{tot}}$ and D boundary conditions on $C := P \times \{0\} \times \mathfrak{g}^*[-1]$. Since we may identify $N^*[1]C$ with $\tilde{M} := P \times \mathfrak{g}[1] \times \mathfrak{g}^*[-1]$, we eventually have the TFT with target $\tilde{M}$ and BV action $\mathsf{S}_{\tilde{F}}^{\mathrm{tot}}$, where $\tilde{F}$ is the Legendre transform of $F$. If $P$ is a domain in $\mathbb{R}^n$, we may now use one coordinate chart and get explicit formulae. This construction turns out to be equivalent to the BRST method. It has a generalization, equivalent to the BV method, when we have a map $J\colon P \to \mathbb{R}^k$ such that $J^{-1}(0)$ is coisotropic.

All the above ideas may in principle be applied to the case when the Poisson manifold $P$ is an infinite-dimensional space of maps (or sections) as in field theory. An $(n+1)$-dimensional field theory on $M \times \mathbb{R}$ is a dynamical system on a symplectic manifold $\mathcal{M}$ of sections on $M$ (or a coisotropic submanifold thereof in gauge theories). The Poisson sigma model version then yields [39] an equivalent $(n+2)$-dimensional field theory on $M \times \Sigma$, with $\Sigma$ the upper half plane.

## References


[1] Alexandrov, M., Kontsevich, M., Schwarz, A., Zaboronsky, O., The geometry of the master equation and topological quantum field theory. *Int. J. Mod. Phys.* **A12** (1997), 1405–1430.

[2] Anselmi, D., Removal of divergences with the Batalin–Vilkovisky formalism. *Class. Quant. Grav.* **11** (1994), 2181–2204.

[3] Anselmi, D., More on the subtraction algorithm. *Class. Quant. Grav.* **12** (1995), 319–350.

[4] Bächtold, M., On the finite dimensional BV formalism. Semesterarbeit 2004, http://www.math.unizh.ch/reports/04_05.pdf.

[5] Batalin, I. A., Vilkovisky, G. A., Relativistic S-matrix of dynamical systems with boson and fermion constraints. *Phys. Lett.* **69 B** (1977), 309–312.

[6] Bates, S., Weinstein, A., *Lectures on the Geometry of Quantization.* Berkeley Mathematics Lecture Notes 8, Amer. Math. Soc., Providence, RI, 1997.

[7] Bayen, F., Flato, M., Frønsdal, C., Lichnerowicz, A., Sternheimer, D., Deformation theory and quantization, I, II. *Ann. Phys.* **111** (1978), 61–110, 111–151.

[8] Becchi, C., Rouet, A., Stora, R., Renormalization of the abelian Higgs–Kibble model. *Comm. Math. Phys.* **42** (1975), 127–162.

[9] Berezin, F. A., *Introduction to Superanalysis* (Edited and with a Foreword by A. A. Kirillov, with an Appendix by V. I. Ogievetsky). Translated from the Russian by J. Niederle and R. Kotecký, ed. by D. Leĭtes, Math. Phys. Appl. Math. 9, D. Reidel Publishing Co., Dordrecht 1987.

[10] Bruguières, A., Cattaneo, A. S., Keller, B., Torossian, C., *Déformation, Quantification, Théorie de Lie*. Panorama et Synthèse, French Mathematical Society, to appear.

[11] Calvo, I., Falceto, F., Poisson reduction and branes in Poisson-Sigma models. *Lett. Math. Phys.* **70** (2004), 231–247.

[12] Calvo, I., Falceto, F., Star products and branes in Poisson-Sigma models. hep-th/0507050.

[13] Cattaneo, A. S., On the BV formalism. Unpublished note; http://www.math.unizh.ch/reports/07_05.pdf.

[14] Cattaneo, A. S., Felder, G., A path integral approach to the Kontsevich quantization formula. *Comm. Math. Phys.* **212** (2000), 591–611.

[15] Cattaneo, A. S., Felder, G., On the AKSZ formulation of the Poisson sigma model. *Lett. Math. Phys.* **56** (2001), 163–179.

[16] Cattaneo, A. S., Felder, G., Coisotropic submanifolds in Poisson geometry and branes in the Poisson sigma model. *Lett. Math. Phys.* **69** (2004), 157–175.

[17] Cattaneo, A. S., Felder, G., Relative formality theorem and quantisation of coisotropic submanifolds. *Adv. Math.*, to appear.

[18] Cattaneo, A. S., Fiorenza, D., Longoni, R., On the Hochschild–Kostant–Rosenberg map for graded manifolds. *Internat. Math. Res. Notices* **62** (2005) 3899–3918.

[19] Cattaneo, A. S., Rossi, C. A., Higher-dimensional $BF$ theories in the Batalin–Vilkovisky formalism: the BV action and generalized Wilson loops. *Comm. Math. Phys.* **221** (2001), 591–657.

[20] Chervov, A., Rybnikov, L., Deformation quantization of submanifolds and reductions via Duflo–Kirillov–Kontsevich map. hep-th/0409005.

[21] Deligne, P., Morgan, J. W., Notes on supersymmetry (following Joseph Bernstein). In *Quantum Fields and Strings: A Course for Mathematicians*. Vol. 1, Amer. Math. Soc., Providence, RI, and Institute for Advanced Study, Princeton, NJ, 1999, 41–97.

[22] Fiorenza, D., An introduction to the Batalin–Vilkovisky formalism. Rencontres mathématiques de Glanon, Edition 2003; math.QA/0402057.

[23] Fradkin, E. S., Fradkina, T. E., Quantization of relativistic systems with boson and fermion first- and second-class constraints. *Phys. Lett.* **72 B** (1978), 343–348.

[24] Gomis, J., Paris, J., Samuel, S., Antibracket, antifields and gauge-theory quantization. *Phys. Rept.* **259** (1995), 1–145.

[25] Henneaux, M., Teitelboim, C., *Quantization of Gauge Systems*. Princeton University Press, Princeton, NJ, 1992.

[26] Hochschild, G., Kostant, B., Rosenberg, A., Differential forms on regular affine algebras. *Trans. Amer. Math. Soc.* **102** (1962), 383–408.

[27] Ikeda, N., Two-dimensional gravity and nonlinear gauge theory. *Ann. Phys.* **235** (1994), 435–464.

[28] Kontsevich, M., Deformation quantization of Poisson manifolds, I. *Lett. Math. Phys.* **66** (2003), 157–216.

[29] Lada, T., Stasheff, J., Introduction to sh Lie algebras for physicists. *Internat. J. Theoret. Phys.* **32** (1993), 1087–1104.

[30] Leĭtes, D. A., Introduction to the theory of supermanifolds. *Uspekhi Mat. Nauk* **35** (1) (1980), 3–57; English transl. *Russian Math. Surveys* **35** (1980), 1–64.

[31] Lyakhovich, S. L., Sharapov, BRST theory without Hamiltonian and Lagrangian. *J. High Energy Phys.* **3** (2005), 011.

[32] Mackenzie, K. C. H., Xu, P., Lie bialgebroids and Poisson groupoids. *Duke Math. J.* **73** (1994), 415–452.

[33] Oh, Y.-G., Park, J.-S., Deformations of coisotropic submanifolds and strong homotopy Lie algebroids. *Invent. Math.* **161** (2005), 287–360.

[34] Roytenberg, D., Courant Algebroids, Derived Brackets and Even Symplectic Supermanifolds. Ph.D. Thesis, Berkeley, 1999; math.DG/9910078.

[35] Roytenberg, D., On the structure of graded symplectic supermanifolds and Courant algebroids. In *Quantization, Poisson Brackets and Beyond* (ed. by Theodore Voronov), Contemp. Math. 315, Amer. Math. Soc., Providence, RI, 2002, 169–185.

[36] Roytenberg, D., AKSZ–BV formalism and Courant algebroid-induced topological field theories. Unpublished.

[37] Schaller, P., Strobl, T., Poisson structure induced (topological) field theories. *Modern Phys. Lett.* A **9** (1994), 3129–3136.

[38] Schwarz, A. S., Geometry of Batalin–Vilkovisky quantization. *Comm. Math. Phys.* **155** (1993), 249–260.

[39] Signori, D., Sottovarietà coisotrope in teorie di campo e quantizzazione. Laurea thesis, Milan University, 2004; http://www.math.unizh.ch/reports/02_05.pdf.

[40] Stasheff, J., Homotopy associativity of H-spaces I, II. *Trans. Amer. Math. Soc.* **108** (1963), 275–312.

[41] Stasheff, J., The intrinsic bracket on the deformation complex of an associative algebra. *J. Pure Appl. Math.* **89** (1993), 231–235.

[42] Tyutin, I. V., Lebedev Institute preprint N39 (1975).

[43] Tulczyjew, W. M., The Legendre transformation. *Ann. Inst. Henri Poincaré* **27** (1977), 101–114.

[44] Varadarajan, V. S., *Supersymmetry for Mathematicians: An Introduction.* Courant Lecture Notes in Math. 11, Amer. Math. Soc., Providence, RI, 2004.

[45] Voronov, T., Higher derived brackets and homotopy algebras. *J. Pure Appl. Algebra* **202** (1–3) (2005), 133–153.

Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

E-mail: alberto.cattaneo@math.unizh.ch

# Matrix ansatz and large deviations of the density in exclusion processes

Bernard Derrida*

**Abstract.** Exclusion processes describe a gas of particles on a lattice with hard core repulsion. When such a lattice gas is maintained in contact with two reservoirs at unequal densities, or driven by an external field, it exhibits a non-equilibrium steady state. In one dimension, a number of properties of this steady state can be calculated exactly using a matrix ansatz. This talk gives a short review on results obtained recently by this matrix ansatz approach.

**Mathematics Subject Classification (2000).** 60K35, 82C26.

**Keywords.** Non-equilibrium statistical mechanics, exclusion processes.

## 1. Introduction

Exclusion processes have been studied for a long time as microscopic models of fluids which satisfy at large scale hydrodynamic equations [2], [13], [27], [28], [36], [39], [47]. They give also some of the simplest examples of non-equilibrium steady state [38], [12], [15], [16], [44], [29]. Here I will try to review a number of recent results on exclusion processes which have been obtained using an exact matrix representation of the weights of microscopic configurations in the non-equilibrium steady state.

One of the simplest cases for which this can be done is the symmetric simple exclusion process defined in Section 2. The matrix ansatz is discussed in Section 3 and the large deviation function of the density is obtained in Section 5 (using an additivity property given in Section 4). Section 6 gives a short review of an alternative approach to calculate this large deviation function, the macroscopic fluctuation theory [3], [4], [5]. Section 7 gives the extension of the matrix ansatz to the asymmetric exclusion process from which one can calculate the phase diagram (Section 8) and the fluctuations of density (Section 9).

*The author thanks T. Bodineau, C. Enaud, M. Evans, V. Hakim, J. L. Lebowitz, V. Pasquier and E. R. Speer. All the results reported in this talk were obtained in a series of works done in collaboration with them.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

## 2. The symmetric simple exclusion process

The symmetric simple exclusion process (SSEP) describes a lattice gas of particles diffusing on a lattice with an exclusion rule which prevents a particle to move to a site already occupied by another particle. Here we consider the one dimensional version with open boundaries. The lattice consists of $L$ sites, each site being either occupied by a single particle or empty. During every infinitesimal time interval $dt$, each particle has a probability $dt$ of jumping to the left if the neighboring site on its left is empty, $dt$ of jumping to the right if the neighboring site on its right is empty. At the two boundaries the dynamics is modified to mimic the coupling with reservoirs of particles: at the left boundary, during each time interval $dt$, a particle is injected on site 1 with probability $\alpha dt$ (if this site is empty) and a particle is removed from site 1 with probability $\gamma dt$ (if this site is occupied). Similarly on site $L$, particles are injected at rate $\delta$ and removed at rate $\beta$.

From the very definition of the SSEP, if $\tau_i = 0$ or 1 is a binary variable indicating whether site $i$ is occupied or empty, one can write the time evolution of the average occupation $\langle \tau_i \rangle$:

$$\begin{aligned} \frac{d\langle \tau_1 \rangle}{dt} &= \alpha - (\alpha + \gamma + 1)\langle \tau_1 \rangle + \langle \tau_2 \rangle, \\ \frac{d\langle \tau_i \rangle}{dt} &= \langle \tau_{i-1} \rangle - 2\langle \tau_i \rangle + \langle \tau_{i+1} \rangle \qquad \text{for } 2 \le i \le L-1, \\ \frac{d\langle \tau_L \rangle}{dt} &= \langle \tau_{L-1} \rangle - (1 + \beta + \delta)\langle \tau_L \rangle + \delta. \end{aligned} \tag{1}$$

The steady state density profile (obtained by writing that $\frac{d\langle \tau_i \rangle}{dt} = 0$) is [20]

$$\langle \tau_i \rangle = \frac{\rho_a \left( L + \frac{1}{\beta+\delta} - i \right) + \rho_b \left( i - 1 + \frac{1}{\alpha+\gamma} \right)}{L + \frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta} - 1} \tag{2}$$

where $\rho_a$ and $\rho_b$ are defined by

$$\rho_a = \frac{\alpha}{\alpha + \gamma}, \quad \rho_b = \frac{\delta}{\beta + \delta}. \tag{3}$$

For a large system size ($L \to \infty$) one can notice that $\langle \tau_1 \rangle \to \rho_a$ and $\langle \tau_L \rangle \to \rho_b$ indicating that $\rho_a$ and $\rho_b$ defined by (3) represent the densities of the left and right reservoirs. One can in fact show [19], [20] that the rates $\alpha, \gamma, \beta, \delta$ do correspond to the left and right boundaries being connected respectively to reservoirs at densities $\rho_a$ and $\rho_b$.

In a similar way one can write down the equations which govern the time evolution of the two point function or higher correlations. For example one finds [46], [23] in

the steady state for $1 \leq i < j \leq L$

$$\begin{aligned}\langle \tau_i \tau_i \rangle_c &\equiv \langle \tau_i \tau_j \rangle - \langle \tau_i \rangle \langle \tau_j \rangle \\ &= -\frac{\left(\frac{1}{\alpha+\gamma} + i - 1\right)\left(\frac{1}{\beta+\delta} + L - j\right)}{\left(\frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta} + L - 1\right)^2 \left(\frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta} + L - 2\right)} (\rho_a - \rho_b)^2.\end{aligned}$$

One can notice that for large $L$, if one introduces macroscopic coordinates $i = Lx$ and $j = Ly$, this becomes

$$\langle \tau_{Lx} \tau_{Ly} \rangle_c = -\frac{x(1-y)}{L} (\rho_a - \rho_b)^2$$

for $x < y$. These weak, but long range, correlations are characteristic of the steady state of non equilibrium systems [47], [23], [41].

The average current in the steady state is given by

$$\bar{j} = \langle \tau_i (1 - \tau_{i+1}) - \tau_{i+1}(1 - \tau_i) \rangle = \langle \tau_i - \tau_{i+1} \rangle = \frac{\rho_a - \rho_b}{L + \frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta} - 1}. \quad (4)$$

This shows that for large $L$, the current $\bar{j} \simeq \frac{\rho_a - \rho_b}{L}$ is proportional to the gradient of the density (with a coefficient of proportionality which is here simply 1) and therefore follows Fick's law.

## 3. The matrix ansatz for the SSEP

For the SSEP, one can write down the steady state equations satisfied by higher and higher correlation functions, but solving these equations becomes quickly inextricable.

The matrix ansatz gives an algebraic way of calculating exactly the weights of all the configurations in the steady state: in [16] it was shown that the probability of a microscopic configuration $\{\tau_1, \tau_2, \ldots, \tau_L\}$ can be written as the matrix element of a product of $L$ matrices

$$\text{Pro}(\{\tau_1, \tau_2, \ldots, \tau_L\}) = \frac{\langle W | X_1 X_2 \ldots X_L | V \rangle}{\langle W | (D + E)^L | V \rangle} \quad (5)$$

where the matrix $X_i$ depends on the occupation $\tau_i$ of site $i$,

$$X_i = \tau_i D + (1 - \tau_i) E, \quad (6)$$

and the matrices $D$ and $E$ satisfy the following algebraic rules:

$$\begin{aligned} DE - ED &= D + E, \\ \langle W | (\alpha E - \gamma D) &= \langle W |, \\ (\beta D - \delta E) | V \rangle &= | V \rangle. \end{aligned} \quad (7)$$

Let us check on a simple example that expression (5) does give the steady state weights: if one chooses the configuration where the first $p$ sites on the left are occupied and the remaining $L-p$ sites on the right are empty, the weight of this configuration is given by

$$\frac{\langle W|D^p E^{L-p}|V\rangle}{\langle W|(D+E)^L|V\rangle}. \tag{8}$$

For (5) to be the weights of all configurations in the steady state, one needs that the rate at which the system enters each configuration and the rate at which the system leaves it should be equal. In the case of the configuration we consider in (8), this means that the following steady state identity should be satisfied:

$$\begin{aligned}(\gamma+1+\delta)\frac{\langle W|D^p E^{L-p}|V\rangle}{\langle W|(D+E)^L|V\rangle} = \alpha\frac{\langle W|ED^{p-1}E^{L-p}|V\rangle}{\langle W|(D+E)^L|V\rangle} \\ + \frac{\langle W|D^{p-1}EDE^{L-p-1}|V\rangle}{\langle W|(D+E)^L|V\rangle} \\ + \beta\frac{\langle W|D^p E^{L-p-1}D|V\rangle}{\langle W|(D+E)^L|V\rangle}.\end{aligned} \tag{9}$$

This equality is easy to check by rewriting (9) as

$$\begin{aligned}\frac{\langle W|(\alpha E-\gamma D)D^{p-1}E^{L-p}|V\rangle}{\langle W|(D+E)^L|V\rangle} - \frac{\langle W|D^{p-1}(DE-ED)E^{L-p-1}|V\rangle}{\langle W|(D+E)^L|V\rangle} \\ + \frac{\langle W|D^p E^{L-p-1}(\beta D-\delta E)|V\rangle}{\langle W|(D+E)^L|V\rangle} = 0\end{aligned} \tag{10}$$

and by using (7). A similar reasoning allows one to prove that the corresponding steady state identity holds for any other configuration.

A priori one should construct the matrices $D$ and $E$ (which might be infinite-dimensional) and the vectors $\langle W|$ and $|V\rangle$ satisfying (7) to calculate the weights of the microscopic configurations. However these weights do not depend on the particular representation chosen and can be calculated directly from (7).

This can be easily seen by using the two matrices $A$ and $B$ defined by

$$\begin{aligned}A &= \beta D-\delta E,\\ B &= \alpha E-\gamma D,\end{aligned} \tag{11}$$

which satisfy

$$AB-BA = (\alpha\beta-\gamma\delta)(D+E) = (\alpha+\gamma)A+(\beta+\delta)B. \tag{12}$$

Each product of $D$'s and $E$'s can be written as a sum of products of $A$'s and $B$'s which can be ordered using (12) by pushing all the $A$'s to the right and all the $B$'s to the left. One gets that way a sum of terms of the form $B^p A^q$, the matrix elements of which can be evaluated easily ($\langle W|B^p A^q|V\rangle = \langle W|V\rangle$) from (7) and (11).

One can calculate that way the average density profile

$$\langle \tau_i \rangle = \frac{\langle W|(D+E)^{i-1} D (D+E)^{L-i}|V\rangle}{\langle W|(D+E)^{L}|V\rangle}$$

as well as all the correlation functions and one can recover that way (2).

One can also show that (equation (3.11) of [20])

$$\frac{\langle W|(D+E)^{L}|V\rangle}{\langle W|V\rangle} = \frac{1}{(\rho_a - \rho_b)^L} \frac{\Gamma(L + \frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta})}{\Gamma(\frac{1}{\alpha+\gamma} + \frac{1}{\beta+\delta})} \tag{13}$$

and using the fact that the average current between sites $i$ and $i+1$ is given by

$$\bar{j} = \frac{\langle W|(D+E)^{i-1}(DE-ED)(D+E)^{L-i-1}|V\rangle}{\langle W|(D+E)^{L}|V\rangle} = \frac{\langle W|(D+E)^{L-1}|V\rangle}{\langle W|(D+E)^{L}|V\rangle}$$

one recovers (4) (of course in the steady state the current does not depend on $i$).

**Remark.** When $\rho_a = \rho_b = r$, i.e. for $\alpha\delta = \beta\gamma$ (see (3)), the two reservoirs are at the same density and the steady state becomes the equilibrium (Gibbs state) of the lattice gas at this density $r$. In this case, the weights of the configurations are those of a Bernoulli measure at density $r$, that is

$$\mathrm{Pro}(\{\tau_1, \tau_2, \ldots, \tau_L\}) = \prod_{i=1}^{L} [r\tau_i + (1-r)(1-\tau_i)] \tag{14}$$

as steady state identities such as (9) can be checked directly for $r = \alpha/(\alpha+\gamma) = \delta/(\beta+\delta)$. All steady state properties can also be recovered by making all the calculations with the matrices (5), (7) for $\rho_a \neq \rho_b$ and by taking the limit $\rho_a \to \rho_b$ in the final expressions, as all the expectations, for a lattice of finite size $L$, are rational functions of $\rho_a$ and $\rho_b$.

## 4. Additivity

As in (5) the weight of each configuration is written as the matrix element of a product of $L$ matrices, one can try to insert at a position $L_1$ a complete basis in order to relate the properties of a lattice of $L$ sites to those of two subsystems of sizes $L_1$ and $L - L_1$.

To do so let us define the following left and right eigenvectors $\langle \rho_a, a|$ and $|\rho_b, b\rangle$ of the operators $\rho_a E - (1-\rho_a)D$ and $(1-\rho_b)D - \rho_b E$:

$$\begin{aligned} \langle \rho_a, a| \, [\rho_a E - (1-\rho_a)D] &= a \langle \rho_a, a|, \\ [(1-\rho_b)D - \rho_b E] \, |\rho_b, b\rangle &= b |\rho_b, b\rangle. \end{aligned} \tag{15}$$

It is easy to see, using the definition (3), that the vectors $\langle W|$ and $|V\rangle$ which appear in (7) are given by

$$\begin{aligned} \langle W| &= \langle \rho_a, (\alpha+\gamma)^{-1}|, \\ |V\rangle &= |\rho_b, (\beta+\delta)^{-1}\rangle. \end{aligned} \tag{16}$$

It is then possible to show, using simply the fact (7) that $DE - ED = D + E$ and the definition of the eigenvectors (15), that (for $\rho_b < \rho_a$)

$$\begin{aligned} &\frac{\langle \rho_a, a|Y_1 Y_2|\rho_b, b\rangle}{\langle \rho_a, a|\rho_b, b\rangle} \\ &= \oint_{\rho_b<|\rho|<\rho_a} \frac{d\rho}{2i\pi} \frac{(\rho_a-\rho_b)^{a+b}}{(\rho_a-\rho)^{a+b}(\rho-\rho_b)} \frac{\langle \rho_a, a|Y_1|\rho, b\rangle}{\langle \rho_a, a|\rho, b\rangle} \frac{\langle \rho, 1-b|Y_2|\rho_b, b\rangle}{\langle \rho, 1-b|\rho_b, b\rangle} \end{aligned} \tag{17}$$

where $Y_1$ and $Y_2$ are arbitrary polynomials of matrices $D$ and $E$. (To prove (17) it is sufficient to establish it when $Y_1$ and $Y_2$ are both of the form $E^n D^{n'}$ as any polynomial can be reduced to a sum of such terms by the relation $DE - ED = D + E$. One can also, and this is easier, prove (17) for $Y_1$ of the form $[\rho_a E - (1-\rho_a)D]^n [D+E]^{n'}$ and $Y_2$ of the form $[D+E]^{n''}[(1-\rho_b)D - \rho_b E]^{n'''}$ and show using $DE - ED = D + E$ that any polynomial $Y_1$ or $Y_2$ can be reduced to a finite sum of such terms).

## 5. Large deviation function of density profiles

If one divides a chain of $L$ sites into $n$ boxes of linear size $l$ (one has of course $n = L/l$ such boxes), one can try to determine the probability of finding a certain density profile $\{\rho_1, \rho_2, \ldots, \rho_n\}$, i.e. the probability of seeing $l\rho_1$ particles in the first box, $l\rho_2$ particles in the second box, $\ldots l\rho_n$ in the $n$th box. For large $L$ one expects the following $L$ dependence of this probability

$$\mathrm{Pro}_L(\rho_1, \ldots, \rho_n|\rho_a, \rho_b) \sim \exp[-L\mathcal{F}_n(\rho_1, \rho_2, \ldots, \rho_n|\rho_a, \rho_b)] \tag{18}$$

where $\mathcal{F}_n$ is a large deviation function. If one defines a reduced coordinate $x$ by

$$i = Lx \tag{19}$$

and if one takes the limit $l \to \infty$ with $l \ll L$ so that the number of boxes becomes infinite, one can define a functional $\mathcal{F}$ for an arbitrary density profile $\rho(x)$

$$\mathrm{Pro}_L(\{\rho(x)\}) \sim \exp[-L\mathcal{F}(\{\rho(x)\}|\rho_a, \rho_b)]. \tag{20}$$

For the SSEP (in one dimension), the functional $\mathcal{F}(\rho(x)|\rho_a, \rho_b)$ is given by the following exact expressions:

*At equilibrium,* i.e. for $\rho_a = \rho_b = r$

$$\mathcal{F}(\{\rho(x)\}|r, r) = \int_0^1 B(\rho(x), r)dx \tag{21}$$

where

$$B(\rho, r) = (1-\rho)\log\frac{1-\rho}{1-r} + \rho\log\frac{\rho}{r}. \tag{22}$$

This can be derived easily. When $\rho_a = \rho_b = r$, the steady state is a Bernoulli measure (14) where all the sites are occupied independently with probability $r$. Therefore if one divides a chain of length $L$ into $L/l$ intervals of length $l$, one has

$$\mathrm{Pro}_L(\rho_1, \ldots, \rho_n|r, r) = \prod_i^{L/l} \frac{l!}{[l\rho_i]!\ [l(1-\rho_i)]!}\, r^{l\rho_i}\ (1-r)^{l(1-\rho_i)} \tag{23}$$

and using Stirling's formula one gets (21), (22).

*For the non-equilibrium case,* i.e. for $\rho_a \neq \rho_b$, it was shown in [19], [4], [20] that

$$\mathcal{F}(\{\rho(x)\}|\rho_a, \rho_b) = \int_0^1 dx\ \left[B(\rho(x), F(x)) + \log\frac{F'(x)}{\rho_b - \rho_a}\right] \tag{24}$$

where the function $F(x)$ is the monotone solution of the differential equation

$$\rho(x) = F + \frac{F(1-F)F''}{F'^2} \tag{25}$$

satisfying the boundary conditions $F(0) = \rho_a$ and $F(1) = \rho_b$.

This expression shows that $\mathcal{F}$ is a *non-local* functional of the density profile $\rho(x)$ as $F(x)$ depends (in a non-linear way) on the profile $\rho(y)$ at all points $y$. For example if the difference $\rho_a - \rho_b$ is small, one can expand $\mathcal{F}$ and obtain an expression where the non-local character of the functional is clearly visible

$$\begin{aligned}
&\mathcal{F}(\{\rho(x)\}|\rho_a, \rho_b) \\
&\quad = \int_0^1 dx\ B(\rho(x), \bar\rho(x)) \\
&\qquad + \frac{(\rho_a - \rho_b)^2}{[\rho_a(1-\rho_a)]^2}\int_0^1 dx \int_x^1 dy\, x(1-y)(\rho(x) - \bar\rho(x))(\rho(y) - \bar\rho(y)) \\
&\qquad + O(\rho_a - \rho_b)^3.
\end{aligned}$$

Here $\bar\rho(x)$ is the most likely profile given by

$$\bar\rho(x) = (1-x)\rho_a + x\rho_b. \tag{26}$$

It would be too long to reproduce here the full derivation of (24), (25) from the matrix ansatz [19], [20]. The idea is to decompose the chain into $L/l$ boxes of $l$ sites

and to sum the weights given by the matrix ansatz (5), (7) over all the microscopic configurations for which the number of particles is $l\rho_1$ in the first box, $l\rho_2$ in the second box,..., $l\rho_n$ in the $n$th box.

A rather easy way to derive (24), (25) is to write (we do it here in the particular case where $a+b=1$, i.e. $\frac{1}{\alpha+\gamma}+\frac{1}{\beta+\delta}=1$, and $\rho_a<\rho_b$) from (17) and (13)

$$P_{nl}(\rho_1,\rho_2,\ldots,\rho_n|\rho_a\rho_b)=\frac{(kl)!((n-k)l)!}{(nl)!}\oint_{\rho_b<|\rho|<\rho_a}\frac{d\rho}{2i\pi}\times\frac{(\rho_a-\rho_b)^{nl+1}}{(\rho_a-\rho)^{kl+1}(\rho-\rho_b)^{(n-k)l+1}}\times P_{kl}(\rho_1,\ldots,\rho_n|\rho_a,\rho)P_{(n-k)l}(\rho_{k+1},\ldots,\rho_n|\rho,\rho_b). \tag{27}$$

Note that in (27) the density $\rho$ has become a complex variable. This is not a difficulty as all the weighs (and therefore the probabilities which appear in (27)) are rational functions of $\rho_a$ and $\rho_b$.

For large $nl$, if one writes $k=nx$, by evaluating (27) at the saddle point one gets

$$\begin{aligned}\mathcal{F}_n(\rho_1,\rho_2,\ldots,\rho_n|\rho_a,\rho_b)&=\max_{\rho_b<F<\rho_a}x\mathcal{F}_k(\rho_1,\ldots,\rho_k|\rho_a,F)\\&\quad+(1-x)\mathcal{F}_{n-k}(\rho_{k+1},\ldots,\rho_n|F,\rho_b)\\&\quad+x\log\left(\frac{\rho_a-F}{x}\right)+(1-x)\log\left(\frac{F-\rho_b}{1-x}\right)-\log(\rho_a-\rho_b).\end{aligned} \tag{28}$$

(Note that to estimate (27) by a saddle point method, one should find the value of $\rho$ which maximizes the integrand over the contour. As the contour is perpendicular to the real axis at their crossing point, this becomes a minimum when $\rho$ varies along the real axis).

If one repeats the same procedure $n$ times, one gets

$$\mathcal{F}_n(\rho_1,\rho_2,\ldots,\rho_n|\rho_a,\rho_b)=\max_{\rho_b=F_0<F_1<\cdots<F_n=\rho_a}\frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_1(\rho_i|F_{i-1},F_i)+\log\left(\frac{(F_{i-1}-F_i)n}{\rho_a-\rho_b}\right). \tag{29}$$

For large $n$, as $F_i$ is monotone, the difference $F_{i-1}-F_i$ is small for almost all $i$ and one can replace $\mathcal{F}_1(\rho_i|F_{i-1},F_i)$ by its equilibrium value $\mathcal{F}_1(\rho_i|F_i,F_i)=B(\rho_i,F_i)$. Therefore (29) becomes (24) in the limit $n\to\infty$, with (25) being the equation satisfied by the optimal $F(x)$.

## 6. The macroscopic fluctuation theory

Bertini, De Sole, Gabrielli, Jona-Lasinio and Landim [3], [4], [5] have developped a different and more general theory to calculate this large deviation functional which

can be summarized as follows: one starts from the expression of the probability $Q(\{\rho(x,s),\, j(x,s)\})$ of observing a certain time dependent macroscopic density profile $\rho(x,s)$ and current profile $j(x,t)$ over a time interval $0 \leq s \leq L^2 t$

$$Q(\{\rho(x,s),\, j(x,s)\}) \sim \max_{\rho(x,s)} \exp\left\{-L \int_{-\infty}^{t} ds \int_0^1 dx \frac{[j + D(\rho)\frac{d\rho}{dx}]^2}{2\sigma(\rho)}\right\} \tag{30}$$

where the current $j(x,s)$ is related to the density profile $\rho(x,s)$ by the conservation law

$$\frac{d\rho(x,s)}{ds} = -\frac{dj(x,s)}{dx} \tag{31}$$

and the functions $D(\rho)$ and $\sigma(\rho)$ are characteristic of the diffusive system studied [9], [10].

Then to calculate the probability of observing a certain density profile $\rho(x)$ in the steady state, one has to find out how this fluctuation is produced. For large $L$, one has to find the optimal path $\rho(x,s)$ for $-\infty < s < t$ in the space of profiles which goes from the typical profile $\bar{\rho}(x)$ to the desired profile $\rho(x)$ and

$$\mathrm{Pro}_L(\{\rho(x)\}) \sim \max_{\rho(x,s)} Q(\{\rho(x,s),\, j(x,s)\}) \tag{32}$$

where the optimal path $\rho(x,s)$ satisfies

$$\rho(x,-\infty) = \bar{\rho}(x),$$
$$\rho(x,t) = \rho(x).$$

Finding this optimal path is usually a hard problem, and so far it has not been possible to find the explicit expression of the functional $\mathcal{F}$ for general $D(\rho)$ and $\sigma(\rho)$. For the SSEP [4], where $D(\rho) = 1$ and $\sigma(\rho) = 2\rho(1-\rho)$, this approach allows one nevertheless to derive (24), (25). It also leads to the same expression of $\mathcal{F}$ as found by the matrix approach [24] in the weakly asymmetric exclusion process and allowed one to calculate the large deviation function $\mathcal{F}$ for the KPM model [35], [7] for which no matrix approach or alternative derivation has been used so far.

The macroscopic fluctuation theory has also been successfully used recently to calculate the fluctuations and the large deviations of the current through diffusive systems [6], [9], [10], [33].

## 7. The matrix approach for the asymmetric exclusion process

The matrix ansatz of Section 3 (which gives the weights of the microscopic configurations in the steady state) has been generalized to describe the steady state of several other systems [1], [8], [11], [14], [18], [25], [30], [31], [32], [37], [40], [42], [43], [45], with of course modified algebraic rules for the matrices the vectors $\langle W|$ and $|V\rangle$.

For example for the asymmetric exclusion process (ASEP), for which the definition is the same as the SSEP of Section 2, except that particles jump at rate 1 to their right and at rate $q \neq 1$ to their left (it the target site is empty), one can show [16], [8], [42], [43] that in this case too, the weights are still given by (5) with the algebra (7) replaced by

$$DE - qED = D + E, \tag{33}$$

$$\langle W|(\alpha E - \gamma D) = \langle W|, \tag{34}$$

$$(\beta D - \delta E)|V\rangle = |V\rangle. \tag{35}$$

One should notice that for the ASEP, the direct approach of calculating the steady state properties by writing the time evolution does not work. Indeed (1) becomes

$$\begin{aligned}
\frac{d\langle\tau_1\rangle}{dt} &= \alpha - (\alpha+\gamma+1)\langle\tau_1\rangle + q\langle\tau_2\rangle + (1-q)\langle\tau_1\tau_2\rangle, \\
\frac{d\langle\tau_i\rangle}{dt} &= \langle\tau_{i-1}\rangle - (1+q)\langle\tau_i\rangle + q\langle\tau_{i+1}\rangle - (1-q)(\langle\tau_{i-1}\tau_i\rangle - \langle\tau_i\tau_{i+1}\rangle), \\
\frac{d\langle\tau_L\rangle}{dt} &= \langle\tau_{L-1}\rangle - (q+\beta+\delta)\langle\tau_L\rangle + \delta - (1-q)\langle\tau_{L-1}\tau_L\rangle,
\end{aligned} \tag{36}$$

and the equations which determine the one-point functions are no longer closed. Therefore all the correlation functions have to be determined at the same time and this is what the matrix ansatz does.

The large deviation function $\mathcal{F}$ of the density defined by (20) has been calculated for the ASEP [21], [22], [24] by an extension of the approach sketched in Sections 4 and 5.

## 8. The phase diagram of the totally asymmetric exclusion process

The last two sections (8 and 9) present two results which can be obtained in the totally asymmetric case (TASEP), i.e. for $q = 0$ (in the particular case where particles are injected only at the left boundary and removed only at the right boundary, i.e. when the input rates $\gamma = \delta = 0$). In this case the algebra (33) becomes

$$DE = D + E, \tag{37}$$

$$\langle W|\alpha E = \langle W|, \tag{38}$$

$$\beta D|V\rangle = |V\rangle. \tag{39}$$

As for the SSEP the average current is still given in terms of the vectors $\langle W|$, $V\rangle$ and of the matrices $D$ and $E$ by

$$\bar{j} = \frac{\langle W|(D+E)^{L-1}|V\rangle}{\langle W|(D+E)^{L}|V\rangle} \tag{40}$$

However as the algebraic rules have changed, the expression of the current is different for the SSEP and the ASEP. From the relation $DE = D + E$ it is easy to prove by recurrence that

$$DF(E) = F(1) + E\,\frac{F(E) - F(1)}{E - 1}$$

for any polynomial $F(E)$ and

$$(D + E)^N = \sum_{p=1}^{N} \frac{p(2N - 1 - p)!}{N!(N - p)!}(E^p + E^{p-1}D + \cdots + D^p).$$

Using the fact that

$$\frac{\langle W|E^m D^n|V\rangle}{\langle W|V\rangle} = \frac{1}{\alpha^m}\frac{1}{\beta^n},$$

one gets [16]

$$\frac{\langle W|(D + E)^N|V\rangle}{\langle W|V\rangle} = \sum_{p=1}^{N} \frac{p(2N - 1 - p)!}{N!(N - p)!}\,\frac{\dfrac{1}{\alpha^{p+1}} - \dfrac{1}{\beta^{p+1}}}{\dfrac{1}{\alpha} - \dfrac{1}{\beta}}. \tag{41}$$

For large $N$ this sum is dominated either by $p \sim 1$, or $p \sim N$ depending one the values of $\alpha$ and $\beta$ and one obtains

$$\frac{\langle W|(D + E)^N|V\rangle}{\langle W|V\rangle} \sim \begin{cases} 4^N & \text{if } \alpha > \frac{1}{2} \text{ and } \beta > \frac{1}{2}, \\ [\beta(1 - \beta)]^{-N} & \text{if } \beta < \alpha \text{ and } \beta < \frac{1}{2}, \\ [\alpha(1 - \alpha)]^{-N} & \text{if } \beta > \alpha \text{ and } \alpha < \frac{1}{2}. \end{cases} \tag{42}$$

This leads to three different expressions of the current (40) for large $L$ corresponding to the three different phases:

- the low density phase ($\beta > \alpha$ and $\alpha < \frac{1}{2}$) where $\bar{j} = \alpha(1 - \alpha)$
- the high density phase ($\alpha > \beta$ and $\beta < \frac{1}{2}$) where $\bar{j} = \beta(1 - \beta)$
- the maximal current phase ($\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$) where $\bar{j} = \frac{1}{4}$

which is the exact phase diagram of the TASEP [38], [15], [16], [44]. The existence of phase transitions [26], [34] in these driven lattice gases is one of the most striking properties of non-equilibrium systems, as it is well known that one dimensional systems at equilibrium with short range interactions cannot exhibit phase transitions.

## 9. Correlation functions in the TASEP and Brownian excursions

For the TASEP, in the maximal current phase ($\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$) one can show [17], using the matrix ansatz, that the correlation function of the occupations of $k$ sites at positions $i_1 = Lx_1$, $i_2 = Lx_2, \ldots, i_k = Lx_k$ with $(Lx_1 < Lx_2 < \cdots < Lx_k)$ are given for large $L$ by

$$\left\langle \left(\tau_{Lx_1} - \frac{1}{2}\right) \ldots \left(\tau_{Lx_k} - \frac{1}{2}\right) \right\rangle = \frac{1}{2^k} \frac{1}{L^{k/2}} \frac{d^k}{dx_1 \ldots dx_k} \langle y_1 \ldots y_k \rangle, \qquad (43)$$

where $y(x)$ is a Brownian excursion between 0 and 1 (a Brownian excursion is a Brownian path constrained to $y(x) > 0$ for $0 < x < 1$ with the boundaries $y(0) = y(1)$). The probability $P(y_1 \ldots y_k; x_1 \ldots x_k)$ of finding the Brownian excursion at positions $y_1 \ldots y_k$ for $0 < x_1 < \cdots < x_k < 1$ is

$$P(y_1 \ldots y_k; x_1 \ldots x_k) = \frac{h_{x_1}(y_1)\, g_{x_2 - x_1}(y_1, y_2) \ldots g_{x_k - x_{k-1}}(y_{k-1}, y_k)\, h_{1-x_k}(y_k)}{\sqrt{\pi}},$$

where $h_x$ and $g_x$ are defined by

$$\begin{cases} h_x(y) = \frac{2y}{x^{3/2}} e^{-y^2/x}, \\ g_x(y, y') = \frac{1}{\sqrt{\pi x}} (e^{-(y-y')^2/x} - e^{-(y+y')^2/x}). \end{cases}$$

One can derive easily (43) in the particular case $\alpha = \beta = 1$ using a representation of (37) which consists of two infinite dimensional bidiagonal matrices

$$D = \sum_{n \geq 1} |n\rangle\langle n| + |n\rangle\langle n+1| = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 1 & 0 & \cdots \\ & & & & \ddots & \ddots & \end{pmatrix}$$

and

$$E = \sum_{n \geq 1} |n\rangle\langle n+1| + |n\rangle\langle n| = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 1 & 0 & 0 & \cdots \\ & & & \ddots & \ddots & & \end{pmatrix}$$

with

$$\langle W| = \langle 1| = (1, 0, 0 \ldots),$$

$$\langle V| = \langle 1| = (1, 0, 0 \ldots).$$

With this representation one can write $\langle W|(D+E)^L|V\rangle$ as a sum over a set $\mathcal{M}_L$ of one dimensional random walks $w$ of $L$ steps which never come back to the origin. Each walk $w$ is defined by a sequence $(n_i(w))$ of $L-1$ heights ($n_i(w) \geq 1$) (with $n_0(w) = n_L(w) = 1$ at the boundaries and the constraint $|n_{i+1} - n_i| \leq 1$). One then has

$$\langle W|(D+E)^L|V\rangle = \sum_{w\in\mathcal{M}_L} \Omega(w),$$

where

$$\Omega(w) = \prod_{i=1}^{L} v(n_{i-1}, n_i) \quad \text{with } v(n,n') = \begin{cases} 2 & \text{if } |n-n'| = 0, \\ 1 & \text{if } |n-n'| = 1, \end{cases}$$

using the fact that $v(n,n') = \langle n|D+E|n'\rangle$ since $D+E$ can be written as

$$D+E = \begin{pmatrix} 2 & 1 & & (0) \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ (0) & & 1 & 2 \end{pmatrix}.$$

Then from the matrix expression one gets $\langle \tau_i \rangle$ and $\langle \tau_i \tau_j \rangle$:

$$\left\langle \left(\tau_{i_1} - \frac{1}{2}\right) \dots \left(\tau_{i_k} - \frac{1}{2}\right) \right\rangle = \frac{1}{2^k} \sum_w \nu(w)(n_{i_1} - n_{i_1-1}) \dots (n_{i_k} - n_{i_k-1}), \quad (44)$$

where $\nu(w)$ is the probability of the walk $w$ induced by the weights $\Omega$:

$$\nu(w) = \frac{\Omega(w)}{\sum'_w \Omega(w')}.$$

The expression (44) is the discrete version of (43). The result (43) can be extended [17] to arbitrary values of $\alpha$ and $\beta$ in the maximal current phase (i.e. for $\alpha > 1/2$ and $\beta > 1/2$).

From this link between the density fluctuations and Brownian excursions, one can show that, for a TASEP of $L$ sites, the number $N$ of particles between sites $Lx_1$ and $Lx_2$, has non-Gaussian fluctuations in the maximal current phase: if one defines the reduced density

$$\mu = \frac{N - L(x_2 - x_1)/2}{\sqrt{L}}$$

one can show [17] that for large $L$

$$P(\mu) = \int_0^\infty dy_1 \int_0^\infty dy_2 \, \frac{1}{\sqrt{2\pi(x_2 - x_1)}} \, \exp\left(-\frac{(2\mu + y_1 - y_2)^2}{x_2 - x_1}\right). \quad (45)$$

According to numerical simulations [17] this distribution (properly rescaled) of the fluctuations of the density remains valid for more general driven systems in their maximal current phase. Of course proving it in a more general case is an interesting open question.

## References


[1] Alcaraz , F. C., Dasmahapatra, S., Rittenberg, V., N-species stochastic models with boundaries and quadratic algebras. *J. Phys. A* **31** (1998), 845–878.

[2] Andjel , E. D., Bramson, M. D., Liggett, T. M., Shocks in the asymmetric exclusion process. *Probab. Theory Related Fields* **78** (1988), 231–247.

[3] Bertini, L., De Sole, A., Gabrielli, D., Jona-Lasinio, G., Landim, C., Fluctuations in stationary non equilibrium states of irreversible processes. *Phys. Rev. Lett.* **87** (2001), 040601.

[4] Bertini, L., De Sole, A., Gabrielli, D., Jona-Lasinio, G., Landim, C., Macroscopic fluctuation theory for stationary non equilibrium states. *J. Statist. Phys.* **107** (2002), 635–675.

[5] Bertini, L., De Sole, A., Gabrielli, D., Jona-Lasinio, G., Landim, C., Minimum dissipation principle in stationary non equilibrium states. *J. Statist. Phys.* **116** (2004), 831–841.

[6] Bertini, L., De Sole, A., Gabrielli, D., Jona-Lasinio, G., Landim, C., Current fluctuations in stochastic lattice gases. *Phys. Rev. Lett.* **94** (2005), 030601.

[7] Bertini, L., Gabrielli, D., Lebowitz, J. L., Large Deviations for a Stochastic Model of Heat Flow. *J. Statist. Phys.* **121** (2005), 843–885.

[8] Blythe, R. A., Evans, M. R., Colaiori, F., Essler, F. H. L., Exact solution of a partially asymmetric exclusion model using a deformed oscillator algebra. *J. Phys. A* **33** (2000), 2313–2332.

[9] Bodineau, T., Derrida, B., Current fluctuations in non-equilibrium diffusive systems: an additivity principle. *Phys. Rev. Lett.* **92** (2004), 180601.

[10] Bodineau, T., Derrida, B., Distribution of current in nonequilibrium diffusive systems and phase transitions. *Phys. Rev. E* **72** (2005), 066110.

[11] Boutillier, C., François, P., Mallick, K., Mallick, S., A matrix ansatz for the diffusion of an impurity in the asymmetric exclusion process. *J. Phys. A* **35** (2002), 9703–9730.

[12] Chowdhury, D., Santen, L., Schadschneider, A., Statistical physics of vehicular traffic and some related systems. *Phys. Rep.* **329** (2000), 199–329.

[13] Demasi, A., Presutti, E., Scacciatelli, E., The weakly asymmetric simple exclusion process. *Ann. Inst. H. Poincaré Probab. Statist.* **25** (1989), 1–38.

[14] Derrida, B., An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Phys. Rep.* **301** (1998), 65–83.

[15] Derrida, B., Domany, E., Mukamel, D., An exact solution of a one-dimensional asymmetric exclusion model with open boundaries. *J. Statist. Phys.* **69** (1992), 667–687.

[16] Derrida, B., Evans, M. R., Hakim, V., Pasquier, V., Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *J. Phys. A* **26** (1993), 1493–1517.

[17] Derrida, B., Enaud, C., Lebowitz, J. L., The asymmetric exclusion process and Brownian excursions. *J. Statist. Phys.* **115** (2004), 365–382.

[18] Derrida, B., Janowsky, S. A., Lebowitz, J. L., Speer, E. R., Exact solution of the totally asymmetric simple exclusion process - shock profiles. *J. Statist. Phys.* **73** (1993), 813–842.

[19] Derrida, B., Lebowitz, J. L., Speer, E. R., Free energy functional for nonequilibrium systems: an exactly solvable case. *Phys. Rev. Lett.* **87** (2001), 150601.

[20] Derrida, B., Lebowitz, J. L., Speer, E. R., Large deviation of the density profile in the steady state of the open symmetric simple exclusion process. *J. Statist. Phys.* **107** (2002), 599–634.

[21] Derrida, B., Lebowitz, J. L., Speer, E. R., Exact free energy functional for a driven diffusive open stationary nonequilibrium system. *Phys. Rev. Lett.* **89** (2002), 030601.

[22] Derrida, B., Lebowitz, J. L., Speer, E. R., Exact large deviation functional of a stationary open driven diffusive system: the asymmetric exclusion process. *J. Statist. Phys.* **110** (2003), 775–810.

[23] Derrida, B., Lebowitz, J. L., Speer, E. R., Entropy of open lattice systems. Preprint 2005, *J. Statist. Phys.*, submitted.

[24] Enaud, C., Derrida, B., Large deviation functional of the weakly asymmetric exclusion process. *J. Statist. Phys.* **114** (2004), 537–562.

[25] Essler, F. H. L., Rittenberg, V., Representations of the quadratic algebra and partially asymmetric diffusion with open boundaries. *J. Phys. A* **29** (1996), 3375–3407.

[26] Evans, M. R., Phase transitions in one-dimensional nonequilibrium systems. *Braz. J. Phys.* **30** (2000), 42–57.

[27] Ferrari, P. A., Shock fluctuations in asymmetric simple exclusion. *Probab. Theory Related Fields* **91** (1992), 81–101.

[28] Ferrari, P. A., Kipnis, C., Saada, E., Microscopic structure of traveling waves in the asymmetric simple exclusion process. *Ann. Probab.* **19** (1991), 226–244.

[29] Hinrichsen, H., Non-equilibrium critical phenomena and phase transitions into absorbing states. *Adv. in Phys.* **49** (2000), 815–958.

[30] Hinrichsen, H., Sandow, S., Peschel, I., On matrix product ground states for reaction-diffusion models. *J. Phys. A* **29** (1996), 2643–2649.

[31] Isaev, A. P., Pyatov, P. N., Rittenberg, V., Diffusion algebras. *J. Phys. A* **34** (2001), 5815–5834.

[32] Jafarpour, F. H., Matrix product states of three families of one-dimensional interacting particle systems. *Physica A* **339** (2004), 369–384.

[33] Jordan, A. N., Sukhorukov, E. V., Pilgram, S., Fluctuation statistics in networks: a stochastic path integral approach. *J. Math. Phys.* **45** (2004), 4386.

[34] Kafri, Y., Levine, E., Mukamel, D., Schütz, G. M., Torok, J., Criterion for phase separation in one-dimensional driven systems. *Phys. Rev. Lett.* **89** (2002), 035702.

[35] Kipnis, C., Marchioro, C., Presutti, E., Heat-flow in an exactly solvable model. *J. Statist. Phys.* **27** (1982), 65–74.

[36] Kipnis, C., Olla, S., Varadhan, S. R. S., Hydrodynamics and large deviations for simple exclusion processes. *Comm. Pure Appl. Math.* **42** (1989), 115–137.

[37] Krebs, K., Sandow, S., Matrix product eigenstates for one-dimensional stochastic models and quantum spin chains. *J. Phys. A* **30** (1997), 3165–3173.

[38] Krug, J., Boundary-induced phase-transitions in driven diffusive systems. *Phys. Rev. Lett.* **67** (1991), 1882–1885.

[39] Liggett, T. M., *Stochastic interacting systems: contact, voter and exclusion processes*. Grundlehren Math. Wiss. 324, Springer-Verlag, Berlin 1999.

[40] Mallick, K., Sandow, S., Finite-dimensional representations of the quadratic algebra: Applications to the exclusion process. *J. Phys. A* **30** (1997), 4513–4526.

[41] Ortiz de árate, J. M., Sengers, J. V., On the physical origin of long-ranged fluctuations in fluids in thermal nonequilibrium states. *J. Statist. Phys.* **115** (2004), 1341–1359.

[42] Sandow, S., Partially asymmetric exclusion process with open boundaries. *Phys. Rev. E* **50** (1994), 2660–2667.

[43] Sasamoto, T., One-dimensional partially asymmetric simple exclusion process with open boundaries: orthogonal polynomials approach. *J. Phys. A* **32** (1999), 7109-7131.

[44] Schütz, G. M., Domany, E., Phase-transitions in an exactly soluble one-dimensional exclusion process. *J. Statist. Phys.* **72** (1993), 277–296.

[45] Speer, E. R., Finite-dimensional representations of a shock algebra. *J. Statist. Phys.* **89**, (1997), 169–175.

[46] Spohn, H., Long range correlations for stochastic lattice gases in a non-equilibrium steady state. *J. Phys. A* **16** (1983), 4275–4291.

[47] Spohn, H., *Large scale dynamics of interacting particles*. Texts and Monographs in Physics, Springer-Verlag, Heidelberg 1991.

Laboratoire de Physique Statistique, Ecole Normale Supérieure, 24, rue Lhomond,
75231 Paris Cedex 05, France
E-mail: derrida@lps.ens.fr

# Correlation functions of the $XXZ$ Heisenberg spin chain: Bethe ansatz approach

Jean Michel Maillet*

**Abstract.** We review recent progress in the computation of correlation functions of the $XXZ$ spin-1/2 chain. We describe both finite and infinite chain results. Long distance asymptotic behavior is discussed. Our method is based on the resolution of the quantum inverse scattering problem in the framework of the algebraic Bethe ansatz.

**Mathematics Subject Classification (2000).** Primary 82B20, 82B23, 82C23, 81R12, 81R50, 81U15, 81U40; Secondary 20G42, 20G45, 37K10, 37K15.

**Keywords.** Quantum integrable models, algebraic Bethe ansatz, correlation functions.

## 1. Introduction

The main challenging problem in the theory of quantum integrable models [1], [2], [3], [4], [5], [6] besides computing their spectrum is to obtain exact and manageable representations for their correlation functions. This issue is of great importance not only from theoretical and mathematical view points but also for applications to relevant physical situations. Although several important advances have been obtained over the years, we are still looking for a general method that could give a systematic solution to this problem. The purpose of this article is to give a review of an approach to this problem elaborated in [7], [8], [9], [10] and in [11], [12], [13], [14], together with a brief account of the more recent progress obtained in [15], [16], [17], [18], [19].

In our search for a general method to compute correlation functions of quantum integrable models our strategy was to consider a simple but representative model where it is possible to develop new tools to solve this problem. Such an archetype of quantum integrable lattice models is provided by the $XXZ$ spin-$\frac{1}{2}$ Heisenberg [20] chain in a magnetic field. Indeed, Heisenberg spin chains play a prominent role in the theory of quantum integrable models: they were the first models for which Bethe ansatz [21], [22], [23], [24], [25] was invented and successfully applied to compute their spectrum. This method has been later used and generalized to solve a large variety of integrable models ranging from statistical mechanics to quantum field theories (see [1], [2], [3], [4], [5]).

*The author would like to thank N. Kitanine, N. Slavnov and V. Terras for their longstanding collaboration on the topics presented in this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The $XXZ$ spin-$\frac{1}{2}$ Heisenberg chain in a magnetic field is a quantum interacting model defined on a one-dimensional lattice with Hamiltonian

$$H = H^{(0)} - hS_z; \tag{1}$$

$$H^{(0)} = \sum_{m=1}^{M} \left\{ \sigma_m^x \sigma_{m+1}^x + \sigma_m^y \sigma_{m+1}^y + \Delta(\sigma_m^z \sigma_{m+1}^z - 1) \right\}, \tag{2}$$

$$S_z = \frac{1}{2} \sum_{m=1}^{M} \sigma_m^z, \quad [H^{(0)}, S_z] = 0. \tag{3}$$

Here $\Delta$ is the anisotropy parameter, $h$ denotes the magnetic field, and $\sigma_m^{x,y,z}$ are the local spin operators (in the spin-$\frac{1}{2}$ representation) associated with each site $m$ of the chain. The quantum space of states is $\mathcal{H} = \otimes_{m=1}^{M} \mathcal{H}_m$, where $\mathcal{H}_m \sim \mathbb{C}^2$ is called local quantum space, with $\dim \mathcal{H} = 2^M$. The operators $\sigma_m^{x,y,z}$ act as the corresponding Pauli matrices in the space $\mathcal{H}_m$ and as the identity operator elsewhere. For simplicity, the length of the chain $M$ is chosen to be even and we assume periodic boundary conditions. Since the simultaneous reversal of all spins is equivalent to a change of sign of the magnetic field, it is enough to consider the case $h \geq 0$.

The first task to solve such a model is to describe the spectrum of its Hamiltonian (1). The method to compute eigenvectors and associated energy levels of the Heisenberg spin chains goes back to H. Bethe in 1931 [21], [22], [23], [25] and is known as the Bethe ansatz. An algebraic version of it has been invented in the late 70s by Faddeev, Sklyanin and Taktajan [26], [27].

The second problem is to compute matrix elements of spin operators $\sigma_m^{x,y,z}$ between two eigenvectors of $H$ and then all correlation functions of spin operators: at zero temperature they reduce to the average value of products of spin operators in the lowest energy level state (the ground state). Let us denote by $| \psi_g \rangle$ the normalized ground state vector. Let $E_m^{\epsilon'_m, \epsilon_m}$ be the elementary operators acting at site $m$ as the $2 \times 2$ matrices $E_{lk}^{\epsilon', \epsilon} = \delta_{l,\epsilon'} \delta_{k,\epsilon}$. Any $n$-point correlation function can be reconstructed as a sum of the following elementary blocks:

$$F_m(\{\epsilon_j, \epsilon'_j\},\ h) = \langle \psi_g | \prod_{j=1}^{m} E_j^{\epsilon'_j, \epsilon_j} | \psi_g \rangle. \tag{4}$$

The knowledge of such correlation functions was for a long time restricted to the free fermion point $\Delta = 0$, a case for which nevertheless tremendous works have been necessary to obtain full answers [28], [29], [30], [31], [32], [33]. For generic $\Delta$, in the thermodynamic limit, at zero temperature, and for zero magnetic field, multiple integral representation of the above elementary blocks of the correlation functions have been obtained from the $q$-vertex operator approach (also using corner transfer matrix technique) in the massive regime $\Delta \geq 1$ in 1992 [34] and conjectured in 1996 [35] for the massless regime $-1 \leq \Delta \leq 1$ (see also [6]).

These results together with their extension to non-zero magnetic field have been obtained in 1999 [8], [9] using the algebraic Bethe ansatz framework [26], [27], [4] and the actual resolution of the so-called quantum inverse scattering problem [8], [10]. This method allows for the computation of the matrix elements of the local spin operators and the above elementary blocks of the correlation functions for the finite chain. Hence, thermodynamic limit can be considered separately. Moreover, time or temperature dependent correlation functions can also be computed [15], [16], [36] using such techniques. Let us mention also recent advances using $q$-KZ equations [37], [38].

This article is meant to be a rather brief review of the problem of correlation functions. More detailed account of the results sketched here together with their proofs can be found in the original articles [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] and in [18], [19]. This article is organized as follows. The space of states of the Heisenberg spin chain will be described in the next section. It includes a brief introduction to the algebraic Bethe ansatz and to various tools of importance in the computation of correlation functions, like in particular the solution of the quantum inverse scattering problem and the determinant representations of the scalar products of states. Section 3 is devoted to the correlation functions of the finite chain. Correlation functions in the thermodynamic limit are studied in Section 4. In Section 5 we describe several exact and asymptotic results together with some open problems. Conclusions and some perspectives are given in the last section.

## 2. The space of states: algebraic Bethe ansatz

The space of states is of dimension $2^M$. As can be observed from the definition of the Hamiltonian in (1), the construction of its eigenvectors is rather non trivial. The purpose of this section is to briefly explain the basics of the knowledge of the space of states in the framework of the algebraic Bethe ansatz, leading in particular to the determination of the spectrum of (1).

**2.1. Algebraic Bethe ansatz.** The algebraic Bethe ansatz originated from the fusion of the original (coordinate) Bethe ansatz and of the inverse scattering method in its Hamiltonian formulation [26], [27], [4]. At the root of the algebraic Bethe ansatz method is the construction of the quantum monodromy matrix. In the case of the $XXZ$ chain (1) the monodromy matrix is a $2 \times 2$ matrix,

$$T(\lambda) = \begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & D(\lambda) \end{pmatrix}, \tag{5}$$

with operator-valued entries $A$, $B$, $C$ and $D$ which depend on a complex parameter $\lambda$ (spectral parameter) and act in the quantum space of states $\mathcal{H}$ of the chain. One of the main property of these operators is that the trace of $T$, namely $A + D$, commutes with the Hamiltonian $H$, while operators $B$ and $C$ can be used as creation operators

of respectively eigenvectors and dual eigenvectors of $A + D$ and hence of $H$ itself. The monodromy matrix is defined as the following ordered product:

$$T(\lambda) = L_M(\lambda) \dots L_2(\lambda) L_1(\lambda), \tag{6}$$

where $L_n(\lambda)$ denotes the quantum $L$-operator at the site $n$ of the chain,

$$L_n(\lambda) = \begin{pmatrix} \sinh(\lambda + \frac{\eta}{2}\sigma_n^z) & \sinh\eta\,\sigma_n^- \\ \sinh\eta\,\sigma_n^+ & \sinh(\lambda - \frac{\eta}{2}\sigma_n^z) \end{pmatrix}. \tag{7}$$

The parameter $\eta$ is related to the anisotropy parameter as $\Delta = \cosh\eta$. It follows from this definition that the monodromy matrix is an highly non local operator in terms of the local spin operators $\sigma_n^{x,y,z}$. However, the commutation relations between the operators $A$, $B$, $C$, $D$ can be computed in a simple way. They are given by the quantum $R$-matrix

$$R(\lambda, \mu) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & b(\lambda,\mu) & c(\lambda,\mu) & 0 \\ 0 & c(\lambda,\mu) & b(\lambda,\mu) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{8}$$

where

$$b(\lambda, \mu) = \frac{\sinh(\lambda - \mu)}{\sinh(\lambda - \mu + \eta)}, \quad c(\lambda, \mu) = \frac{\sinh(\eta)}{\sinh(\lambda - \mu + \eta)}. \tag{9}$$

The $R$-matrix is a linear operator in the tensor product $V_1 \otimes V_2$, where each $V_i$ is isomorphic to $\mathbb{C}^2$, and depends generically on two spectral parameters $\lambda_1$ and $\lambda_2$ associated to these two vector spaces. It is denoted by $R_{12}(\lambda_1, \lambda_2)$. Such an $R$-matrix satisfies the Yang–Baxter equation,

$$R_{12}(\lambda_1, \lambda_2)\ R_{13}(\lambda_1, \lambda_3)\ R_{23}(\lambda_2, \lambda_3) = R_{23}(\lambda_2, \lambda_3)\ R_{13}(\lambda_1, \lambda_3)\ R_{12}(\lambda_1, \lambda_2). \tag{10}$$

It gives the following commutation relations among the operators entries of the monodromy matrix:

$$R_{12}(\lambda, \mu)\ T_1(\lambda)\ T_2(\mu) = T_2(\mu)\ T_1(\lambda)\ R_{12}(\lambda, \mu) \tag{11}$$

with the tensor notations $T_1(\lambda) = T(\lambda) \otimes \mathrm{Id}$ and $T_2(\mu) = \mathrm{Id} \otimes T(\mu)$. These commutation relations imply in particular that the transfer matrices, defined as

$$\mathcal{T}(\lambda) = \mathrm{tr}\, T(\lambda) = A(\lambda) + D(\lambda), \tag{12}$$

commute for different values of the spectral parameter $[\mathcal{T}(\lambda), \mathcal{T}(\mu)] = 0$ and also with $S_z$, $[\mathcal{T}(\lambda), S_z] = 0$. The Hamiltonian (2) at $h = 0$ is related to $\mathcal{T}(\lambda)$ by the 'trace identity'

$$H^{(0)} = 2\sinh\eta\ \frac{d\mathcal{T}(\lambda)}{d\lambda}\mathcal{T}^{-1}(\lambda)\big|_{\lambda=\frac{\eta}{2}} - 2M\cosh\eta. \tag{13}$$

Therefore, the spectrum of the Hamiltonian (1) is given by the common eigenvectors of the transfer matrices and of $S_z$.

For technical reasons, it is actually convenient to introduce a slightly more general object, the twisted transfer matrix

$$\mathcal{T}_\kappa(\lambda) = A(\lambda) + \kappa D(\lambda), \tag{14}$$

where $\kappa$ is a complex parameter. The particular case of $\mathcal{T}_\kappa(\lambda)$ at $\kappa = 1$ corresponds to the usual (untwisted) transfer matrix $\mathcal{T}(\lambda)$. It will be also convenient to consider an inhomogeneous version of the $XXZ$ chain for which

$$T_{1\dots M}(\lambda; \xi_1, \dots, \xi_M) = L_M(\lambda - \xi_M + \eta/2) \dots L_1(\lambda - \xi_1 + \eta/2). \tag{15}$$

Here, $\xi_1, \dots, \xi_M$ are complex parameters (inhomogeneity parameters) attached to each site of the lattice. The homogeneous model (1) corresponds to the case where $\xi_j = \eta/2$ for $j = 1, \dots, M$.

In the framework of algebraic Bethe ansatz, an arbitrary quantum state can be obtained from the vectors generated by multiple action of operators $B(\lambda)$ on the reference vector $|\,0\,\rangle$ with all spins up (respectively by multiple action of operators $C(\lambda)$ on the dual reference vector $\langle\,0\,|$),

$$|\,\psi\,\rangle = \prod_{j=1}^{N} B(\lambda_j)|\,0\,\rangle, \quad \langle\,\psi\,| = \langle\,0\,|\prod_{j=1}^{N} C(\lambda_j), \quad N = 0, 1, \dots, M. \tag{16}$$

**2.2. Description of the spectrum.** Let us consider here the subspace $\mathcal{H}^{(M/2-N)}$ of the space of states $\mathcal{H}$ with a fixed number $N$ of spins down. In this subspace, the eigenvectors $|\,\psi_\kappa(\{\lambda\})\,\rangle$ (respectively $\langle\,\psi_\kappa(\{\lambda\})\,|$) of the twisted transfer matrix $\mathcal{T}_\kappa(\mu)$ can be constructed in the form (16), where the parameters $\lambda_1, \dots, \lambda_N$ satisfy the system of twisted Bethe equations

$$\mathcal{Y}_\kappa(\lambda_j|\{\lambda\}) = 0, \quad j = 1, \dots, N. \tag{17}$$

Here, the function $\mathcal{Y}_\kappa$ is defined as

$$\mathcal{Y}_\kappa(\mu|\{\lambda\}) = a(\mu)\prod_{k=1}^{N}\sinh(\lambda_k - \mu + \eta) + \kappa\, d(\mu)\prod_{k=1}^{N}\sinh(\lambda_k - \mu - \eta), \tag{18}$$

and $a(\lambda)$, $d(\lambda)$ are the eigenvalues of the operators $A(\lambda)$ and $D(\lambda)$ on the reference state $|\,0\,\rangle$. In the normalization (7) and for the inhomogeneous model (15) we have

$$a(\lambda) = \prod_{a=1}^{M}\sinh(\lambda - \xi_a + \eta), \quad d(\lambda) = \prod_{a=1}^{M}\sinh(\lambda - \xi_a). \tag{19}$$

The corresponding eigenvalue of $\mathcal{T}_\kappa(\mu)$ on $|\,\psi_\kappa(\{\lambda\})\,\rangle$ (or on a dual eigenvector) is

$$\tau_\kappa(\mu|\{\lambda\}) = a(\mu)\prod_{k=1}^{N}\frac{\sinh(\lambda_k-\mu+\eta)}{\sinh(\lambda_k-\mu)} + \kappa\, d(\mu)\prod_{k=1}^{N}\frac{\sinh(\mu-\lambda_k+\eta)}{\sinh(\mu-\lambda_k)}. \quad (20)$$

The solutions of the system of twisted Bethe equations (17) have been analyzed in [39]. In general, not all of these solutions correspond to eigenvectors of $\mathcal{T}_\kappa(\mu)$.

**Definition 2.1.** A solution $\{\lambda\}$ of the system (17) is called *admissible* if

$$d(\lambda_j)\prod_{\substack{k=1\\k\neq j}}^{N}\sinh(\lambda_j-\lambda_k+\eta)\neq 0, \quad j=1,\ldots,N, \quad (21)$$

and *un-admissible* otherwise. A solution is called *off-diagonal* if the corresponding parameters $\lambda_1,\ldots,\lambda_N$ are pairwise distinct, and *diagonal* otherwise.

One of the main result of [39] is that, for generic parameters $\kappa$ and $\{\xi\}$, the set of the eigenvectors corresponding to the admissible off-diagonal solutions of the system of twisted Bethe equations (17) form a basis in the subspace $\mathcal{H}^{(M/2-N)}$. It has been proven in [16] that this result is still valid in the homogeneous case $\xi_j=\eta/2$, $j=1,\ldots,N$, at least if $\kappa$ is in a punctured vicinity of the origin (i.e. $0<|\kappa|<\kappa_0$ for $\kappa_0$ small enough). Note however that, for specific values of $\kappa$ and $\{\xi\}$, the basis of the eigenvectors in $\mathcal{H}^{(M/2-N)}$ may include some states corresponding to un-admissible solutions of (17) (in particular in the homogeneous limit at $\kappa=1$).

At $\kappa=1$, it follows from the trace identity (13) that the eigenvectors of the transfer matrix coincide, in the homogeneous limit, with the ones of the Hamiltonian (1). The corresponding eigenvalues in the case of zero magnetic field can be obtained from (13), (20):

$$H^{(0)}\,|\,\psi(\{\lambda\})\,\rangle = \Big(\sum_{j=1}^{N}E(\lambda_j)\Big)\cdot|\,\psi(\{\lambda\})\,\rangle, \quad (22)$$

where the (bare) one-particle energy $E(\lambda)$ is equal to

$$E(\lambda)=\frac{2\sinh^2\eta}{\sinh(\lambda+\frac{\eta}{2})\sinh(\lambda-\frac{\eta}{2})}. \quad (23)$$

**2.3. Drinfel'd twist and $F$-basis.** As already noted, the operators $A$, $B$, $C$, $D$ are highly non local in terms of local spin operators. There exists however an interesting description of these operators by means of a change of basis of the space of states. In particular, this basis will provide a direct access to the scalar products of states. The root of this new basis is provided by the notion of Drinfel'd twist [40] associated to the $R$-matrix of the $XXZ$ chain. It leads to the notion of factorizing $F$-matrices. To be essentially self-contained we briefly recall here their main properties and refer to [7] for more details and proofs.

**Definition 2.2.** For inhomogeneity parameters $\xi_j$ in generic positions and for any integer $n$ one can associate to any element $\sigma$ of the symmetric group $S_n$ on $n$ elements a unique $R$-matrix $R^{\sigma}_{1\ldots n}(\xi_1, \ldots, \xi_n)$, denoted for simplicity $R^{\sigma}_{1\ldots n}$, constructed as an ordered product (depending on $\sigma$) of the elementary $R$-matrices $R_{ij}(\xi_i, \xi_j)$.

We have the following property for an arbitrary integer $n$.

**Proposition 2.1.**

$$R^{\sigma}_{1\ldots n}\; T_{1\ldots n}(\lambda; \xi_1, \ldots, \xi_n) = T_{\sigma(1)\ldots\sigma(n)}(\lambda; \xi_{\sigma(1)}, \ldots, \xi_{\sigma(n)})\; R^{\sigma}_{1\ldots n}. \tag{24}$$

We can now define the notion of a factorizing $F$-matrix:

**Definition 2.3.** A factorizing $F$-matrix associated to a given elementary $R$-matrix is an invertible matrix $F_{1\ldots n}(\xi_1, \ldots, \xi_n)$, defined for an arbitrary integer $n$, satisfying the following relation for any element $\sigma$ of $S_n$:

$$F_{\sigma(1)\ldots\sigma(n)}(\xi_{\sigma(1)}, \ldots, \xi_{\sigma(n)})\; R^{\sigma}_{1\ldots n}(\xi_1, \ldots, \xi_n) = F_{1\ldots n}(\xi_1, \ldots, \xi_n). \tag{25}$$

In other words, such an $F$-matrix factorizes the corresponding $R$-matrix for arbitrary integers $n$. Taking into account the fact that the parameters $\xi_n$ are in one to one correspondence with the vector spaces $\mathcal{H}_n$, we can adopt simplified notations such that

$$F_{1\ldots n}(\xi_1, \ldots, \xi_n) = F_{1\ldots n},$$
$$F_{\sigma(1)\ldots\sigma(n)}(\xi_{\sigma(1)}, \ldots, \xi_{\sigma(n)}) = F_{\sigma(1)\ldots\sigma(n)}.$$

**Theorem 2.1** ([7]). *For the $XXZ$ model with inhomogeneity parameters $\xi_n$ in generic positions there exists a factorizing, triangular $F$-matrix. It is constructed explicitly from the $R$-matrix.*

This matrix has two important properties:

**Proposition 2.2** ([7]). *In the $F$-basis, the monodromy matrix $\widetilde{T}$,*

$$\widetilde{T}_{1\ldots M}(\lambda; \xi_1, \ldots, \xi_M) \;=\; F_{1\ldots M} T_{1\ldots M}(\lambda; \xi_1, \ldots, \xi_M)\; F^{-1}_{1\ldots M}, \tag{26}$$

*is totally symmetric under any simultaneous permutations of the lattice sites $i$ and of the corresponding inhomogeneity parameters $\xi_i$.*

The second property gives the explicit expressions of the monodromy matrix in the $F$-basis. For the $XXZ$-$\frac{1}{2}$ model, the quantum monodromy operator is a $2 \times 2$ matrix with entries $A, B, C, D$ which are obtained as sums of $2^{M-1}$ operators, which themselves are products of $M$ local spin operators on the quantum chain. As an example, the $B$ operator is given as

$$B_{1\ldots M}(\lambda) = \sum_{i=1}^{N} \sigma_i^- \;\Omega_i \;+\; \sum_{i\neq j\neq k} \sigma_i^- \;(\sigma_j^- \;\sigma_k^+)\; \Omega_{ijk} \;+\; \text{higher terms}, \tag{27}$$

where the matrices $\Omega_i$, $\Omega_{ijk}$, are diagonal operators acting respectively on all sites but $i$, on all sites but $i, j, k$, and the higher order terms involve more and more exchange spin terms like $\sigma_j^- \sigma_k^+$. It means that the $B$ operator returns one spin somewhere on the chain, this operation being however dressed non-locally and with non-diagonal operators by multiple exchange terms of the type $\sigma_j^- \sigma_k^+$.

So, whereas these formulas in the original basis are quite involved, their expressions in the $F$-basis simplify drastically:

**Proposition 2.3** ([7]). *The operators D, B and C in the F-basis are given by the formulas*

$$\widetilde{D}_{1\ldots M}(\lambda;\xi_1,\ldots,\xi_M) = \bigotimes_{i=1}^{M} \begin{pmatrix} b(\lambda,\xi_i) & 0 \\ 0 & 1 \end{pmatrix}_{[i]}, \tag{28}$$

$$\widetilde{B}_{1\ldots M}(\lambda) = \sum_{i=1}^{M} \sigma_i^- \, c(\lambda,\xi_i) \bigotimes_{j\neq i} \begin{pmatrix} b(\lambda,\xi_j) & 0 \\ 0 & b^{-1}(\xi_j,\xi_i) \end{pmatrix}_{[j]}, \tag{29}$$

$$\widetilde{C}_{1\ldots M}(\lambda) = \sum_{i=1}^{M} \sigma_i^+ \, c(\lambda,\xi_i) \bigotimes_{j\neq i} \begin{pmatrix} b(\lambda,\xi_j)\, b^{-1}(\xi_i,\xi_j) & 0 \\ 0 & 1 \end{pmatrix}_{[j]}, \tag{30}$$

*and the operator $\widetilde{A}$ can be obtained from quantum determinant relations.*

We wish first to stress that while the operators $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$ satisfy the same quadratic commutation relations as $A, B, C, D$, they are completely symmetric under simultaneous exchange of the inhomogeneity parameters and of the spaces $\mathcal{H}_n$. It really means that the factorizing $F$-matrices we have constructed solve the combinatorial problem induced by the non-trivial action of the permutation group $S_M$ given by the $R$-matrix. In the $F$-basis the action of the permutation group on the operators $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$ is trivial.

Further, it can be shown that the pseudo-vacuum vector is left invariant, namely, it is an eigenvector of the total $F$-matrix with eigenvalue 1; in particular, the algebraic Bethe ansatz can be carried out also in the $F$-basis. Hence, a direct computation of Bethe eigenvectors and of their scalar products in this $F$-basis is made possible, while it was a priori very involved in the original basis. There, only commutation relations between the operators $A, B, C, D$ can be used, leading (see [5]) to very intricate sums over partitions.

**2.4. Solution of the quantum inverse problem.** The very simple expressions of the monodromy matrix operators entries $D$, $B$, $C$ in the $F$-basis suggests that any local operator $E_j^{\epsilon'_j,\epsilon_j}$, acting in a local quantum space $\mathcal{H}_j$ at site $j$, can be expressed in terms of the entries of the monodromy matrix. This is the so-called quantum inverse scattering problem. The solution to this problem was found in [8], [10]:

**Theorem 2.2.**

$$E_j^{\epsilon_j',\epsilon_j} = \prod_{\alpha=1}^{j-1} \mathcal{T}(\xi_\alpha) \cdot T_{\epsilon_j,\epsilon_j'}(\xi_j) \cdot \prod_{\alpha=1}^{j} \mathcal{T}^{-1}(\xi_\alpha). \tag{31}$$

The proof of this theorem is elementary (see [8], [10]) and hence it can be obtained for a large class of lattice integrable models. It relies essentially on the property that the $R$-matrix $R(\lambda, \mu)$ reduces to the permutation operator for $\lambda = \mu$. An immediate consequence of this theorem is that the operators $A$, $B$, $C$, and $D$ generate the space of all operators acting in $\mathcal{H}$.

**2.5. Scalar products.** We give here the expressions for the scalar product of an eigenvector of the twisted transfer matrix with any arbitrary state of the form (16). These scalar products can be expressed as determinant of rather simple matrices. The root of all these determinants is in fact the determinant representation for the partition function of the 6-vertex model with domain wall boundary conditions [41]. Let us first define, for arbitrary positive integers $n, n'$ ($n \leq n'$) and arbitrary sets of variables $\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_n$ and $\nu_1, \ldots, \nu_{n'}$ such that $\{\lambda\} \subset \{\nu\}$, the $n \times n$ matrix $\Omega_\kappa(\{\lambda\}, \{\mu\}|\{\nu\})$ as

$$\begin{aligned}(\Omega_\kappa)_{jk}(\{\lambda\}, \{\mu\}|\{\nu\}) = a(\mu_k)\, t(\lambda_j, \mu_k) \prod_{a=1}^{n'} \sinh(\nu_a - \mu_k + \eta) \\ - \kappa\, d(\mu_k)\, t(\mu_k, \lambda_j) \prod_{a=1}^{n'} \sinh(\nu_a - \mu_k - \eta),\end{aligned} \tag{32}$$

with

$$t(\lambda, \mu) = \frac{\sinh \eta}{\sinh(\lambda - \mu) \sinh(\lambda - \mu + \eta)}. \tag{33}$$

**Proposition 2.4** ([42], [8], [15]). *Let $\{\lambda_1, \ldots, \lambda_N\}$ be a solution of the system of twisted Bethe equations* (17)*, and let $\mu_1, \ldots, \mu_N$ be generic complex numbers. Then,*

$$\begin{aligned}\langle 0 | \prod_{j=1}^{N} C(\mu_j) | \psi_\kappa(\{\lambda\}) \rangle &= \langle \psi_\kappa(\{\lambda\}) | \prod_{j=1}^{N} B(\mu_j) | 0 \rangle \\ &= \frac{\prod_{a=1}^{N} d(\lambda_a) \prod_{a,b=1}^{N} \sinh(\mu_b - \lambda_a)}{\prod_{a>b}^{N} \sinh(\lambda_a - \lambda_b) \sinh(\mu_b - \mu_a)} \cdot \det_N \left( \frac{\partial}{\partial \lambda_j} \tau_\kappa(\mu_k | \{\lambda\}) \right)\end{aligned} \tag{34}$$

$$= \frac{\prod_{a=1}^{N} d(\lambda_a)}{\prod_{a>b}^{N} \sinh(\lambda_a - \lambda_b)\sinh(\mu_b - \mu_a)} \cdot \det_N \Omega_\kappa(\{\lambda\},\{\mu\}|\{\lambda\}). \quad (35)$$

These equations are valid for any arbitrary complex parameter $\kappa$, in particular at $\kappa = 1$. In this case we may omit the subscript $\kappa$ and denote $(\psi, \tau, \mathcal{Y}, \Omega) = (\psi_\kappa, \tau_\kappa, \mathcal{Y}_\kappa, \Omega_\kappa)\big|_{\kappa=1}$. If the sets $\{\lambda\}$ and $\{\mu\}$ are different, the eigenvector $|\, \psi_\kappa(\{\lambda\})\,\rangle$ is orthogonal to the dual eigenvector $\langle\, \psi_\kappa(\{\mu\})\,|$. Otherwise we obtain a formula for the norm of the corresponding vector [43], [44], [8],

$$\langle\, \psi_\kappa(\{\lambda\}) \,|\, \psi_\kappa(\{\lambda\})\,\rangle = \frac{\prod_{a=1}^{N} d(\lambda_a)}{\prod_{\substack{a,b=1\\ a\neq b}}^{N} \sinh(\lambda_a - \lambda_b)} \cdot \det_N \Omega_\kappa(\{\lambda\},\{\lambda\}|\{\lambda\})$$

$$= (-1)^N \frac{\prod_{a=1}^{N} d(\lambda_a)}{\prod_{\substack{a,b=1\\ a\neq b}}^{N} \sinh(\lambda_a - \lambda_b)} \cdot \det_N \left( \frac{\partial}{\partial \lambda_k} \mathcal{Y}_\kappa(\lambda_j|\{\lambda\}) \right).$$

**2.6. Action of operators $A$, $B$, $C$, $D$ on a general state.** An important step of the computation of the correlation function is to express the action of any product of local operators on any Bethe eigenvector. From the solution of the quantum inverse scattering problem, this is given by the successive action of $A$, $B$, $C$, $D$ operators on a vector constructed by action of $C$ operators on the reference vector. Action of $A$, $B$, $C$, $D$ on such a vector are well known (see for example [5]). They can be written in the following form:

$$\langle\, 0 \,|\, \prod_{k=1}^{N} C(\lambda_k)\, A(\lambda_{N+1}) = \sum_{a'=1}^{N+1} a(\lambda_{a'}) \frac{\prod_{k=1}^{N} \sinh(\lambda_k - \lambda_{a'} + \eta)}{\prod_{\substack{k=1\\ k\neq a'}}^{N+1} \sinh(\lambda_k - \lambda_{a'})} \langle\, 0 \,|\, \prod_{\substack{k=1\\ k\neq a'}}^{N+1} C(\lambda_k); \quad (36)$$

$$\langle\, 0 \,|\, \prod_{k=1}^{N} C(\lambda_k)\, D(\lambda_{N+1}) = \sum_{a=1}^{N+1} d(\lambda_a) \frac{\prod_{k=1}^{N} \sinh(\lambda_a - \lambda_k + \eta)}{\prod_{\substack{k=1\\ k\neq a}}^{N+1} \sinh(\lambda_a - \lambda_k)} \langle\, 0 \,|\, \prod_{\substack{k=1\\ k\neq a}}^{N+1} C(\lambda_k). \quad (37)$$

The action of the operator $B(\lambda)$ can be obtained similarly,

$$\begin{aligned}\langle 0 \mid \prod_{k=1}^{N} C(\lambda_k)\, B(\lambda_{N+1}) = \sum_{a=1}^{N+1} d(\lambda_a) \frac{\prod_{k=1}^{N} \sinh(\lambda_a - \lambda_k + \eta)}{\prod_{\substack{k=1 \\ k\neq a}}^{N+1} \sinh(\lambda_a - \lambda_k)} \\ \times \sum_{\substack{a'=1 \\ a'\neq a}}^{N+1} \frac{a(\lambda_{a'})}{\sinh(\lambda_{N+1} - \lambda_{a'} + \eta)} \frac{\prod_{\substack{j=1 \\ j\neq a}}^{N+1} \sinh(\lambda_j - \lambda_{a'} + \eta)}{\prod_{\substack{j=1 \\ j\neq a,a'}}^{N+1} \sinh(\lambda_j - \lambda_{a'})} \langle 0 \mid \prod_{\substack{k=1 \\ k\neq a,a'}}^{N+1} C(\lambda_k),\end{aligned} \tag{38}$$

and the action of $C$ is obvious.

## 3. Correlation functions: finite chain

To compute correlation functions of some product of local operators, the following successive problems have to be addressed: (i) determination of the ground state $\langle\, \psi_g \mid$, (ii) evaluation of the action of the product of the local operators on it, and (iii) computation of the scalar product of the resulting state with $\mid \psi_g \,\rangle$. Using the solution of the quantum inverse scattering problem together with the explicit determinant formulas for the scalar products and the norm of the Bethe state, one sees that matrix elements of local spin operators and correlation functions can be expressed as (multiple) sums of determinants [9]. It should be stressed that this result is purely algebraic and is valid for finite chains of arbitrary length $M$.

**3.1. Matrix elements of local operators.** We begin with the calculation of the one-point functions. These results follow directly from the solution of the quantum inverse scattering problem, the above action of operators $A$, $B$, $C$ and $D$, and the determinant representation of the scalar products. We consider

$$F_N^-(m, \{\mu_j\}, \{\lambda_k\}) = \langle 0 \mid \prod_{j=1}^{N+1} C(\mu_j)\, \sigma_m^- \prod_{k=1}^{N} B(\lambda_k) \mid 0 \,\rangle \tag{39}$$

and

$$F_N^+(m, \{\lambda_k\}, \{\mu_j\}) = \langle 0 \mid \prod_{k=1}^{N} C(\lambda_k)\, \sigma_m^+ \prod_{j=1}^{N+1} B(\mu_j) \mid 0 \,\rangle, \tag{40}$$

where $\{\lambda_k\}_n$ and $\{\mu_j\}_{n+1}$ are solutions of Bethe equations.

**Proposition 3.1.** *For two Bethe states with spectral parameters* $\{\lambda_k\}_N$ *and* $\{\mu_j\}_{N+1}$, *the matrix element of the operator* $\sigma_m^-$ *can be represented as a determinant,*

$$F_N^-(m,\{\mu_j\},\{\lambda_k\}) = \frac{\phi_{m-1}(\{\mu_j\})}{\phi_{m-1}(\{\lambda_k\})} \frac{\prod_{j=1}^{N+1} \sinh(\mu_j - \xi_m + \eta)}{\prod_{k=1}^{N} \sinh(\lambda_k - \xi_m + \eta)} \cdot \frac{\det_{N+1} H^-(m,\{\mu_j\},\{\lambda_k\})}{\prod_{N+1\geq k>j\geq 1} \sinh(\mu_k - \mu_j) \prod_{1\leq\beta<\alpha\leq N} \sinh(\lambda_\beta - \lambda_\alpha)}, \tag{41}$$

$$\phi_m(\{\lambda_k\}) = \prod_{k=1}^{N}\prod_{j=1}^{m} b^{-1}(\lambda_k,\xi_j), \tag{42}$$

*and the* $(N+1)\times(N+1)$ *matrix* $H^-$ *is defined as*

$$H_{ab}^-(m) = \frac{\varphi(\eta)}{\varphi(\mu_a - \lambda_b)}\Big(a(\lambda_b)\prod_{\substack{j=1\\ j\neq a}}^{N+1}\varphi(\mu_j - \lambda_b + \eta) - d(\lambda_b)\prod_{\substack{j=1\\ j\neq a}}^{N+1}\varphi(\mu_j - \lambda_b - \eta)\Big) \tag{43}$$

*for* $b < N+1$, *and*

$$H_{aN+1}^-(m) = \frac{\varphi(\eta)}{\varphi(\mu_a - \xi_m + \eta)\varphi(\mu_a - \xi_m)}. \tag{44}$$

*For the matrix element* $F_N^+(m,\{\lambda_k\},\{\mu_j\})$ *we get*

$$F_N^+(m,\{\lambda_k\},\{\mu_j\}) = \frac{\phi_m(\lambda_k)\,\phi_{m-1}(\lambda_k)}{\phi_{m-1}(\mu_j)\,\phi_m(\mu_j)} F_N^-(m,\{\mu_j\},\{\lambda_k\}). \tag{45}$$

The matrix elements of the operator $\sigma_m^z$ between two Bethe states have been obtained similarly [8].

**3.2. Elementary blocks of correlation functions.** In this section we consider a more general case of correlation functions: the ground state mean value of any product of the local elementary $2\times 2$ matrices $E_{lk}^{\epsilon',\epsilon} = \delta_{l,\epsilon'}\delta_{k,\epsilon}$:

$$F_m(\{\epsilon_j,\epsilon_j'\}) = \frac{\langle\,\psi_g\,|\prod_{j=1}^{m} E_j^{\epsilon_j',\epsilon_j}|\,\psi_g\,\rangle}{\langle\,\psi_g\,|\psi_g\rangle}. \tag{46}$$

An arbitrary $n$-point correlation function can be obtained as a sum of such mean values. Using the solution of the quantum inverse scattering problem, we reduce this

problem to the computation of the ground state mean value of an arbitrary ordered product of monodromy matrix elements,

$$F_m(\{\epsilon_j,\epsilon'_j\}) = \phi_m^{-1}(\{\lambda\})\frac{\langle\,\psi_g\,|T_{\epsilon_1,\epsilon'_1}(\xi_1)\dots T_{\epsilon_m,\epsilon'_m}(\xi_m)|\,\psi_g\,\rangle}{\langle\,\psi_g\,|\psi_g\rangle}. \tag{47}$$

To calculate these mean values we first describe generically the product of the monodromy matrix elements. For that purpose, one should consider the two following sets of indices, $\boldsymbol{\alpha}^+ = \{j : 1 \le j \le m,\ \epsilon_j = 1\}$, $\mathrm{card}(\boldsymbol{\alpha}^+) = s'$, $\max_{j\in\boldsymbol{\alpha}^+}(j) \equiv j'_{\max}$, $\min_{j\in\boldsymbol{\alpha}^+}(j) \equiv j'_{\min}$, and similarly $\boldsymbol{\alpha}^- = \{j : 1 \le j \le m,\ \epsilon'_j = 2\}$, $\mathrm{card}(\boldsymbol{\alpha}^-) = s$, $\max_{j\in\boldsymbol{\alpha}^-}(j) \equiv j_{\max}$, $\min_{j\in\boldsymbol{\alpha}^-}(j) \equiv j_{\min}$. The intersection of these two sets is not empty and corresponds to the operators $B(\xi_j)$. Consider now the action, $\langle\,0\,|\prod_{k=1}^N C(\lambda_k)T_{\epsilon_1,\epsilon'_1}(\lambda_{N+1})\dots T_{\epsilon_m,\epsilon'_m}(\lambda_{N+m})$, applying one by one the formulae (36)–(38). For all the indices $j$ from the sets $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ one obtains a summation on the corresponding indices $a'_j$ (for $j \in \boldsymbol{\alpha}^+$, corresponding to the action of the operators $A(\lambda)$ or $B(\lambda)$) or $a_j$ (for $j \in \boldsymbol{\alpha}^-$, corresponding to the action of the operators $D(\lambda)$ or $B(\lambda)$). As the product of the monodromy matrix elements is ordered these summations are also ordered and the corresponding indices should be taken from the following sets: $\boldsymbol{A}_j = \{b : 1 \le b \le N+m,\ b \ne a_k, a'_k,\ k < j\}$ and $\boldsymbol{A'}_j = \{b : 1 \le b \le N+m,\ b \ne a'_k,\ k < j,\ b \ne a_k,\ k \le j\}$. Thus,

$$\begin{aligned}\langle\,0\,|\prod_{k=1}^{N} C(\lambda_k)\,T_{\epsilon_1,\epsilon'_1}(\lambda_{N+1})\dots T_{\epsilon_m,\epsilon'_m}(\lambda_{N+m})\\ = \sum_{\{a_j,a'_j\}} G_{\{a_j,a'_j\}}(\lambda_1,\dots,\lambda_{N+m})\langle\,0\,|\prod_{b\in\boldsymbol{A}_{m+1}} C(\lambda_b).\end{aligned} \tag{48}$$

The summation is taken over the indices $a_j$ for $j \in \boldsymbol{\alpha}^-$ and $a'_j$ for $j \in \boldsymbol{\alpha}^+$ such that $1 \le a_j \le N+j$, $a_j \in \boldsymbol{A}_j$, $1 \le a'_j \le N+j$, $a'_j \in \boldsymbol{A'}_j$. The functions $G_{\{a_j,a'_j\}}(\lambda_1,\dots\lambda_{N+m})$ can then be easily obtained from the formulae (36)–(38) taking into acount that $\lambda_a = \xi_{a-N}$ for $a > N$:

$$\begin{aligned}G_{\{a_j,a'_j\}}(\lambda_1,\dots,\lambda_{N+m}) = \prod_{j\in\boldsymbol{\alpha}^-} d(\lambda_{a_j})\frac{\prod\limits_{\substack{b=1\\ b\in\boldsymbol{A}_j}}^{N+j-1}\sinh(\lambda_{a_j}-\lambda_b+\eta)}{\prod\limits_{\substack{b=1\\ b\in\boldsymbol{A'}_j}}^{N+j}\sinh(\lambda_{a_j}-\lambda_b)}\\ \times\prod_{j\in\boldsymbol{\alpha}^+} a(\lambda_{a'_j})\frac{\prod\limits_{\substack{b=1\\ b\in\boldsymbol{A'}_j}}^{N+j-1}\sinh(\lambda_b-\lambda_{a'_j}+\eta)}{\prod\limits_{\substack{b=1\\ b\in\boldsymbol{A}_{j+1}}}^{N+j}\sinh(\lambda_b-\lambda_{a'_j})}.\end{aligned} \tag{49}$$

Now to calculate the normalized mean value (47) we apply the determinant representation for the scalar product. It should be mentioned that the number of operators $C(\lambda)$ has to be equal to the number of the operators $B(\lambda)$, as otherwise the mean value is zero, and hence the total number of elements in the sets $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ is $s + s' = m$. Taking into account that in (47), for $b > N$, $\lambda_b = \xi_{b-N}$ one has to consider the following scalar products:

$$\frac{\langle 0 \mid \prod\limits_{b \in \boldsymbol{A}_{m+1}} C(\lambda_b) \prod\limits_{k=1}^{N} B(\lambda_k) \mid 0 \rangle}{\langle 0 \mid \prod\limits_{k=1}^{N} C(\lambda_k) \prod\limits_{k=1}^{N} B(\lambda_k) \mid 0 \rangle},$$

for all the permitted values of $a_j, a'_j$. Finally we obtain

$$F_m(\{\epsilon_j, \epsilon'_j\}) = \frac{1}{\prod\limits_{k<l} \sinh(\xi_k - \xi_l)} \sum_{\{a_j, a'_j\}} H_{\{a_j, a'_j\}}(\lambda_1, \ldots, \lambda_{N+m}), \tag{50}$$

the sum being taken on the same set of indices $a_j$, $a'_j$ as in (48). The functions $H_{\{a_j, a'_j\}}(\{\lambda\})$ can be obtained using (49) and the determinant representations for the scalar products.

**3.3. Two-point functions.** The method presented in the last section is quite straightforward and gives formally the possibility to compute any correlation function. However, it has been developed for the computation of the expectation values of the monomials $T_{a_1 b_1}(\xi_1) \ldots T_{a_m b_m}(\xi_m)$, leading to the evaluation of elementary building blocks, whereas the study of the two-point functions involves big sums of such monomials. Indeed, let us consider for example the correlation function $\langle \sigma_1^z \sigma_{m+1}^z \rangle$. Then, according to the solution of the inverse scattering problem (31), we need to calculate the expectation value

$$\langle \psi(\{\lambda\}) \mid (A - D)(\xi_1) \cdot \prod_{a=2}^{m} \mathcal{T}(\xi_a) \cdot (A - D)(\xi_{m+1}) \cdot \prod_{b=1}^{m+1} \mathcal{T}^{-1}(\xi_b) \mid \psi(\{\lambda\}) \rangle. \tag{51}$$

Since $\mid \psi(\{\lambda\}) \rangle$ is an eigenvector, the action of $\prod_{b=1}^{m+1} \mathcal{T}^{-1}(\xi_b)$ on this state merely produces a numerical factor. However, it is much more complicated to evaluate the action of $\prod_{a=2}^{m} \mathcal{T}(\xi_a)$. Indeed, we have to act first with $(A - D)(\xi_1)$ on $\langle \psi(\{\lambda\}) \mid$ (or with $(A - D)(\xi_{m+1})$ on $\mid \psi(\{\lambda\}) \rangle$), which gives a sum of states which are no longer eigenvectors of the transfer matrix, and on which the multiple action of $\mathcal{T}$ is not simple. In fact, the product $\prod_{a=2}^{m}(A + D)(\xi_a)$ would lead to a sum of $2^{m-1}$ elementary blocks. This is not very convenient, in particular at large distance $m$. Therefore, to obtain manageable expressions for such correlation functions, it is of great importance to develop an alternative and compact way to express the multiple

action of the transfer matrix on arbitrary states or, in other words, to make an effective re-summation of the corresponding sum of the $2^{m-1}$ terms. This can be achieved in the following way:

**Proposition 3.2.** *Let $\kappa$, $x_1, \dots, x_m$ and $\mu_1, \dots, \mu_N$ be generic parameters. Then the action of $\prod_{a=1}^{m} \mathcal{T}_\kappa(x_a)$ on a state of the form $\langle 0 | \prod_{j=1}^{N} C(\mu_j)$ can be formally written as*

$$\langle 0 | \prod_{j=1}^{N} C(\mu_j) \prod_{a=1}^{m} \mathcal{T}_\kappa(x_a) = \frac{1}{N!} \oint\limits_{\Gamma\{x\}\cup\Gamma\{\mu\}} \prod_{j=1}^{N} \frac{dz_j}{2\pi i} \cdot \prod_{a=1}^{m} \tau_\kappa(x_a|\{z\}) \cdot \prod_{a=1}^{N} \frac{1}{\mathcal{Y}_\kappa(z_a|\{z\})} \times \prod_{\substack{j,k=1 \\ j<k}}^{N} \frac{\sinh(z_j - z_k)}{\sinh(\mu_j - \mu_k)} \cdot \det_N \Omega_\kappa(\{z\}, \{\mu\}|\{z\}) \cdot \langle 0 | \prod_{j=1}^{N} C(z_j), \quad (52)$$

*where the integration contour $\Gamma\{x\} \cup \Gamma\{\mu\}$ surrounds the points*[1] *$x_1, \dots, x_m$ and $\mu_1, \dots, \mu_N$ and does not contain any other pole of the integrand.*

One of the simplest applications concerns the generating function of the two-point correlation function of the third components of spin, which is defined as the normalized expectation value $\langle\, Q^\kappa_{l,m} \,\rangle$ of the operator

$$Q^\kappa_{l,m} = \prod_{n=l}^{m} \left( \frac{1+\kappa}{2} + \frac{1-\kappa}{2} \cdot \sigma^z_n \right) = \prod_{j=1}^{l-1} \mathcal{T}(\xi_j) \cdot \prod_{j=l}^{m} \mathcal{T}_\kappa(\xi_j) \cdot \prod_{j=1}^{m} \mathcal{T}^{-1}(\xi_j), \quad (53)$$

where $|\, \psi(\{\lambda\}) \,\rangle$ is an eigenvector of $\mathcal{T}(\mu)$ in the subspace $\mathcal{H}^{(M/2-N)}$. The two-point correlation function of the third components of local spins in the eigenvector $|\, \psi(\{\lambda\}) \,\rangle$ can be obtained in terms of the second 'lattice derivative' and the second derivative with respect to $\kappa$ of the generating function $\langle\, Q^\kappa_{l,m} \,\rangle$ at $\kappa = 1$:

$$\langle\, \sigma^z_l \sigma^z_{l+m} \,\rangle = \langle\, \sigma^z_l \,\rangle + \langle\, \sigma^z_{l+m} \,\rangle - 1 + 2 \frac{\partial^2}{\partial \kappa^2} \langle\, Q^\kappa_{l,l+m} - Q^\kappa_{l,l+m-1} - Q^\kappa_{l+1,l+m} + Q^\kappa_{l+1,l+m-1} \,\rangle\big|_{\kappa=1}. \quad (54)$$

Due to the translational invariance of the correlation functions in the homogeneous model, we will simply consider the expectation value $\langle\, Q^\kappa_{1,m} \,\rangle$. For any given eigenvector, we obtain the following result:

**Theorem 3.1.** *Let $\{\lambda\}$ be an admissible off-diagonal solution of the system of untwisted Bethe equations, and let us consider the corresponding expectation value $\langle\, Q^\kappa_{1,m} \,\rangle$ in*

[1] More precisely, for a set of complex variables $\{\nu_1, \dots, \nu_l\}$, the notation $\Gamma\{\nu\}$ should be understood in the following way: $\Gamma\{\nu\}$ is the boundary of a set of poly-disks $\mathcal{D}_a(r)$ in $\mathbb{C}^N$, i.e. $\Gamma\{\nu\} = \cup_{a=1}^{l} \bar{\mathcal{D}}_a(r)$ with $\bar{\mathcal{D}}_a(r) = \{z \in \mathbb{C}^N : |z_k - \nu_a| = r, \quad k = 1, \dots, N\}$.

*the inhomogeneous finite XXZ chain. Then there exists* $\kappa_0 > 0$ *such that, for* $|\kappa| < \kappa_0$, *the following representations hold:*

$$\langle\, Q^{\kappa}_{1,m}\,\rangle = \frac{1}{N!}\oint\limits_{\Gamma\{\xi\}\cup\Gamma\{\lambda\}} \prod_{j=1}^{N}\frac{dz_j}{2\pi i}\cdot\prod_{a=1}^{m}\frac{\tau_\kappa(\xi_a|\{z\})}{\tau(\xi_a|\{\lambda\})}\cdot\prod_{a=1}^{N}\frac{1}{\mathcal{Y}_\kappa(z_a|\{z\})} \times \det_N \Omega_\kappa(\{z\},\{\lambda\}|\{z\})\cdot\frac{\det_N \Omega(\{\lambda\},\{z\}|\{\lambda\})}{\det_N \Omega(\{\lambda\},\{\lambda\}|\{\lambda\})}. \tag{55}$$

*The integration contours are such that the only singularities of the integrand which contribute to the integral are the points* $\xi_1,\ldots,\xi_m$ *and* $\lambda_1\ldots,\lambda_N$.

From this result we can extract a compact representation for the two-point function of $\sigma^z$ [15]. Similar expressions exists for other correlation functions of the spin operators, and in particular for the time dependent case [15], [16]. Moreover, this multiple contour integral representation permits to relate two very different ways to compute two point correlation functions of the type, $g_{12} = \langle\omega|\theta_1\theta_2|\omega\rangle$, namely,

(i) to compute the action of local operators on the ground state $\theta_1\theta_2|\omega\rangle = |\tilde{\omega}\rangle$ and then to calculate the resulting scalar product $g_{12} = \langle\omega|\tilde{\omega}\rangle$ as was explained in the previous sections;

(ii) to insert a sum over a complete set of states $|\omega_i\rangle$ (for instance, a complete set of eigenvectors of the Hamiltonian) between the local operators $\theta_1$ and $\theta_2$ and to obtain the representation for the correlation function as a sum over matrix elements of local operators,

$$g_{12} = \sum_i \langle\omega|\theta_1|\omega_i\rangle\cdot\langle\omega_i|\theta_2|\omega\rangle. \tag{56}$$

In fact the above representation as multiple contour integrals contains both expansions. Indeed there are two ways to evaluate the corresponding integrals: either to compute the residues in the poles inside $\Gamma$, or to compute the residues in the poles within strips of the width $i\pi$ outside $\Gamma$.

The first way leads to a representation of the correlation function $\langle\sigma^z_1\sigma^z_{m+1}\rangle$ in terms of the previously obtained [11] $m$-multiple sums. Evaluation of the above contour integral in terms of the poles outside the contour $\Gamma$ gives us the expansion (ii) of the correlation function (i.e. an expansion in terms of matrix elements of $\sigma^z$ between the ground state and all excited states). This relation holds also for the time dependent case [15], [16].

## 4. Correlation functions: infinite chain

In the thermodynamic limit $M \to \infty$ and at zero magnetic field, the model exhibits three different regimes depending on the value of $\Delta$ [1]. For $\Delta < -1$, the model is ferromagnetic, for $-1 < \Delta < 1$ the model has a non degenerated anti ferromagnetic

ground state, and no gap in the spectrum (massless regime), while for $\Delta > 1$ the ground state is twice degenerated with a gap in the spectrum (massive regime). In both cases, the ground state has spin zero. Hence the number of parameters $\lambda$ in the ground state vectors is equal to half the size $M$ of the chain. For $M \to \infty$, these parameters will be distributed in some continuous interval according to a density function $\rho$.

**4.1. The thermodynamic limit.** In this limit, the Bethe equations for the ground state, written in their logarithmic form, become a linear integral equation for the density distribution of these $\lambda$'s,

$$\rho_{\mathrm{tot}}(\alpha) + \int_{-\Lambda}^{\Lambda} K(\alpha-\beta)\rho_{\mathrm{tot}}(\beta)\,d\beta = \frac{p'_{0_{\mathrm{tot}}}(\alpha)}{2\pi}, \tag{57}$$

where the new real variables $\alpha$ are defined in terms of general spectral parameters $\lambda$ differently in the two domains. From now on, we only describe the massless regime (see [9] for the other case) $-1 < \Delta < 1$ where $\alpha = \lambda$. The density $\rho$ is defined as the limit of the quantity $\frac{1}{M(\lambda_{j+1}-\lambda_j)}$, and the functions $K(\lambda)$ and $p'_{0_{\mathrm{tot}}}(\lambda)$ are the derivatives with respect to $\lambda$ of the functions $-\frac{\theta(\lambda)}{2\pi}$ and $p_{0_{\mathrm{tot}}}(\lambda)$:

$$\begin{aligned} K(\alpha) &= \frac{\sin 2\zeta}{2\pi\ \sinh(\alpha+i\zeta)\sinh(\alpha-i\zeta)} \\ p'_0(\alpha) &= \frac{\sin\zeta}{\sinh\left(\alpha+i\frac{\zeta}{2}\right)\sinh\left(\alpha-i\frac{\zeta}{2}\right)} \end{aligned} \qquad \text{for } -1<\Delta<1,\ \zeta = i\eta, \tag{58}$$

with $p'_{0_{\mathrm{tot}}}(\alpha) = \frac{1}{M}\sum_{i=1}^{M} p'_0\left(\alpha-\beta_k-i\frac{\zeta}{2}\right)$, where $\beta_k = \xi_k$. The integration limit $\Lambda$ is equal to $+\infty$ for $-1 < \Delta < 1$. The solution for the equation (57) in the homogeneous model where all parameters $\xi_k$ are equal to $\eta/2$, that is the density for the ground state of the Hamiltonian in the thermodynamic limit, is given by the following function [24]:

$$\rho(\alpha) = \frac{1}{2\zeta\cosh(\frac{\pi\alpha}{\zeta})}.$$

For technical convenience, we will also use the function

$$\rho_{\mathrm{tot}}(\alpha) = \frac{1}{M}\sum_{i=1}^{M}\rho\left(\alpha-\beta_k-i\frac{\zeta}{2}\right).$$

It will be also convenient to consider, without any loss of generality, that the inhomogeneity parameters are contained in the region $-\zeta < \mathrm{Im}\beta_j < 0$. Using these results, for any $\mathcal{C}^\infty$ function $f$ ($\pi$-periodic in the domain $\Delta > 1$), sums over all the values of $f$ at the point $\alpha_j$, $1 \le j \le N$, parameterizing the ground state, can be replaced in the thermodynamic limit by an integral:

$$\frac{1}{M}\sum_{j=1}^{N} f(\alpha_j) = \int_{-\Lambda}^{\Lambda} f(\alpha)\rho_{\mathrm{tot}}(\alpha)\,d\alpha + O(M^{-1}).$$

Thus, multiple sums obtained in correlation functions will become multiple integrals. Similarly, it is possible to evaluate the behavior of the determinant formulas for the scalar products and the norm of Bethe vectors (and in particular their ratios) in the limit $M \to \infty$.

**4.2. Elementary blocks.** From the representations as multiple sums of these elementary blocks in the finite chain we can obtain their multiple integral representations in the thermodynamic limit. Let us now consider separately the two regimes of the $XXZ$ model. In the massless regime $\eta = -i\zeta$ is imaginary, the ground state parameters $\lambda$ are real and the limit of integration is infinity $\Lambda = \infty$. In this case we consider the inhomogeneity parameters $\xi_j$ such that $0 > \mathrm{Im}(\xi_j) > -\zeta$. For the correlation functions in the thermodynamic limit one obtains the following result in this regime.

**Proposition 4.1.**

$$F_m(\{\epsilon_j, \epsilon'_j\}) = \prod_{k<l} \frac{\sinh\frac{\pi}{\zeta}(\xi_k - \xi_l)}{\sinh(\xi_k - \xi_l)} \prod_{j=1}^{s'} \int_{-\infty - i\zeta}^{\infty - i\zeta} \frac{d\lambda_j}{2i\zeta} \prod_{j=s'+1}^{m} \int_{-\infty}^{\infty} i\frac{d\lambda_j}{2\zeta}$$

$$\prod_{a=1}^{m}\prod_{k=1}^{m} \frac{1}{\sinh\frac{\pi}{\zeta}(\lambda_a - \xi_k)} \prod_{j\in\boldsymbol{\alpha}^-} \Big(\prod_{k=1}^{j-1} \sinh(\mu_j - \xi_k - i\zeta) \prod_{k=j+1}^{m} \sinh(\mu_j - \xi_k)\Big)$$

$$\prod_{j\in\boldsymbol{\alpha}^+} \Big(\prod_{k=1}^{j-1} \sinh(\mu'_j - \xi_k + i\zeta) \prod_{k=j+1}^{m} \sinh(\mu'_j - \xi_k)\Big) \prod_{a>b} \frac{\sinh\frac{\pi}{\zeta}(\lambda_a - \lambda_b)}{\sinh(\lambda_a - \lambda_b - i\zeta)},$$

*where the parameters of integration are ordered in the following way:* $\{\lambda_1, \ldots \lambda_m\} = \{\mu'_{j'_{\max}}, \ldots, \mu'_{j'_{\min}}, \mu_{j_{\min}}, \ldots, \mu_{j_{\max}}\}$.

The homogeneous limit ($\xi_j = -i\zeta/2$, for all $j$) of the correlation function $F_m(\{\epsilon_j, \epsilon'_j\})$ can then be taken in an obvious way. We have obtained similar representations for the massive regime, and also in the presence of a non-zero magnetic field [9]. For zero magnetic field, these results agree exactly with the ones obtained by Jimbo and Miwa in [35], using in particular $q$-KZ equations. It means that for zero magnetic field, the elementary blocks of correlation functions indeed satisfy $q$-KZ equations. Recently, more algebraic representations of solutions of the $q$-KZ equations have been obtained that correspond to the above correlation functions [37], [38]. From the finite chain representation for the two-point function it is also possible to obtain multiple integral representations for that case as well, in particular for their generating function [11], [13]. They correspond different huge re-summations and symmetrization of the corresponding elementary blocks, as in the finite chain situation [11]. Moreover, the case of time dependent correlation functions as also been obtained [15], [16]. Finally, let us note that at the free fermion point, all the results presented here lead, in a very elementary way, to already know results [12], [17], [19].

## 5. Exact and asymptotic results

**5.1. Exact results at $\Delta = 1/2$.** Up to now, two exact results have been obtained for the case of anisotropy $\Delta = 1/2$: the exact value of the emptiness formation probability for arbitrary distance $m$ [13] and the two point function of the third component of spin [18]. These two results follow from the above multiple integral representations for which, due to the determinant structure of the integrand, the corresponding multiple integrals can be separated and hence explicitly computed for this special value of the anisotropy.

**5.1.1. The emptiness formation probability.** This correlation function $\tau(m)$ (the probability to find in the ground state a ferromagnetic string of length $m$) is defined as the following expectation value:

$$\tau(m) = \langle \psi_g | \prod_{k=1}^{m} \frac{1-\sigma_k^z}{2} | \psi_g \rangle, \tag{59}$$

where $|\psi_g\rangle$ denotes the normalized ground state. In the thermodynamic limit ($M \to \infty$), this quantity can be expressed as a multiple integral with $m$ integrations [34], [35], [6], [8], [9].

**Proposition 5.1.** *For* $\Delta = \cos\zeta$, $0 < \zeta < \pi$, $\tau(m) = \lim_{\xi_1,\dots\xi_m \to -\frac{i\zeta}{2}} \tau(m, \{\xi_j\})$, *where*

$$\tau(m, \{\xi_j\}) = \frac{1}{m!} \int_{-\infty}^{\infty} \frac{Z_m(\{\lambda\}, \{\xi\})}{\prod_{a<b}^{m} \sinh(\xi_a - \xi_b)} \det_m \left( \frac{i}{2\zeta \sinh \frac{\pi}{\zeta}(\lambda_j - \xi_k)} \right) d^m\lambda, \tag{60}$$

$$Z_m(\{\lambda\}, \{\xi\}) = \prod_{a=1}^{m} \prod_{b=1}^{m} \frac{\sinh(\lambda_a - \xi_b)\sinh(\lambda_a - \xi_b - i\zeta)}{\sinh(\lambda_a - \lambda_b - i\zeta)} \cdot \frac{\det_m \left( \frac{-i\sin\zeta}{\sinh(\lambda_j - \xi_k)\sinh(\lambda_j - \xi_k - i\zeta)} \right)}{\prod_{a>b}^{m} \sinh(\xi_a - \xi_b)}. \tag{61}$$

The proof is given in [11]. Due to the determinant structure of the integrand, the integrals can be separated and computed for the special case $\Delta = \frac{1}{2}$ ($\zeta = \pi/3$):

**Proposition 5.2.** *Let* $\xi_k = \varepsilon_k - i\pi/6$ *and* $\varepsilon_{ab} = \varepsilon_a - \varepsilon_b$. *Then we obtain*

$$\tau(m, \{\varepsilon_j\}) = \frac{(-1)^{\frac{m^2-m}{2}}}{2^{m^2}} \prod_{a>b}^{m} \frac{\sinh 3\varepsilon_{ba}}{\sinh \varepsilon_{ba}} \prod_{\substack{a,b=1 \\ a\neq b}}^{m} \frac{1}{\sinh \varepsilon_{ab}} \cdot \det_m \left( \frac{3\sinh \frac{\varepsilon_{jk}}{2}}{\sinh \frac{3\varepsilon_{jk}}{2}} \right) \tag{62}$$

*and*

$$\tau(m) = \left(\frac{1}{2}\right)^{m^2} \prod_{k=0}^{m-1} \frac{(3k+1)!}{(m+k)!}. \tag{63}$$

Observe that the quantity $A_m = \prod_{k=0}^{m-1}(3k+1)!/(m+k)!$ is the number of alternating sign matrices of size $m$. This result was conjectured in [45].

**5.1.2. The two point function of $\sigma^z$.** The two point functions can be obtained, as in the finite chain situation, from a generating function $\langle Q_\kappa(m)\rangle$; in the thermodynamic limit, we use the following multiple integral representation [18]:

$$\begin{aligned}\langle Q_\kappa(m)\rangle = \sum_{n=0}^{m} \frac{\kappa^{m-n}}{n!(m-n)!} \oint_{\Gamma\{-i\zeta/2\}} \frac{d^m z}{(2\pi i)^m} \int_{\mathbb{R}-i\zeta} d^n\lambda \int_{\mathbb{R}} d^{m-n}\lambda \cdot \prod_{j=1}^{m} \frac{\varphi^m(z_j)}{\varphi^m(\lambda_j)} \\ \times \prod_{j=1}^{n} \left\{ t(z_j,\lambda_j) \prod_{k=1}^{m} \frac{\sinh(z_j-\lambda_k-i\zeta)}{\sinh(z_j-z_k-i\zeta)} \right\} \\ \times \prod_{j=n+1}^{m} \left\{ t(\lambda_j,z_j) \prod_{k=1}^{m} \frac{\sinh(\lambda_k-z_j-i\zeta)}{\sinh(z_k-z_j-i\zeta)} \right\} \\ \times \prod_{j=1}^{m}\prod_{k=1}^{m} \frac{\sinh(\lambda_k-z_j-i\zeta)}{\sinh(\lambda_k-\lambda_j-i\zeta)} \cdot \det_m \left( \frac{i}{2\zeta \sinh\frac{\pi}{\zeta}(\lambda-z)} \right).\end{aligned} \tag{64}$$

Here,

$$\Delta = \cos\zeta, \quad t(z,\lambda) = \frac{-i\sin\zeta}{\sinh(z-\lambda)\sinh(z-\lambda-i\zeta)}, \quad \varphi(z) = \frac{\sinh(z-i\frac{\zeta}{2})}{\sinh(z+i\frac{\zeta}{2})}, \tag{65}$$

and the integrals over the variables $z_j$ are taken with respect to a closed contour $\Gamma$ which surrounds the point $-i\zeta/2$ and does not contain any other singularities of the integrand. The equation (64) is valid for the homogeneous $XXZ$ chain with arbitrary $-1 < \Delta < 1$. If we consider the inhomogeneous $XXZ$ model with inhomogeneities $\xi_1, \ldots, \xi_m$, then one should replace in the representation (64) the function $\varphi^m$ in the following way:

$$\varphi^m(z) \to \prod_{b=1}^{m} \frac{\sinh(z-\xi_b-i\zeta)}{\sinh(z-\xi_b)}, \quad \varphi^{-m}(\lambda) \to \prod_{b=1}^{m} \frac{\sinh(\lambda-\xi_b)}{\sinh(\lambda-\xi_b-i\zeta)}. \tag{66}$$

In order to come back to the homogeneous case, one should set $\xi_k = -i\zeta/2$, $k = 1, \ldots, m$ in (66). In the inhomogeneous model, the integration contour $\Gamma$ surrounds the points $\xi_1, \ldots, \xi_m$, and the integrals over $z_j$ are therefore equal to the sum of the residues of the integrand in these simple poles. It turns out that again for the special case $\Delta = \frac{1}{2}$ integrals can be separated and computed to give:

**Proposition 5.3.**

$$\langle Q_\kappa(m)\rangle = \frac{3^m}{2^{m^2}} \prod_{a>b}^{m} \frac{\sinh 3(\xi_a-\xi_b)}{\sinh^3(\xi_a-\xi_b)} \sum_{n=0}^{m} \kappa^{m-n} \sum_{\substack{\{\xi\}=\{\xi_{\gamma_+}\}\cup\{\xi_{\gamma_-}\}\\ |\gamma_+|=n}} \det_m \hat{\Phi}^{(n)}$$

$$\times \prod_{a\in\gamma_+}\prod_{b\in\gamma_-} \frac{\sinh(\xi_b-\xi_a-\frac{i\pi}{3})\sinh(\xi_a-\xi_b)}{\sinh^2(\xi_b-\xi_a+\frac{i\pi}{3})},$$

$$\hat{\Phi}^{(n)}(\{\xi_{\gamma_+}\},\{\xi_{\gamma_-}\}) = \left(\begin{array}{c|c} \Phi(\xi_j-\xi_k) & \Phi(\xi_j-\xi_k-\frac{i\pi}{3}) \\ \hline \Phi(\xi_j-\xi_k+\frac{i\pi}{3}) & \Phi(\xi_j-\xi_k) \end{array}\right),$$

$$\Phi(x) = \frac{\sinh\frac{x}{2}}{\sinh\frac{3x}{2}}.$$

*Here the sum is taken with respect to all partitions of the set $\{\xi\}$ into two disjoint subsets $\{\xi_{\gamma_+}\}\cup\{\xi_{\gamma_-}\}$ of cardinality $n$ and $m-n$ respectively. The first $n$ lines and columns of the matrix $\hat{\Phi}^{(n)}$ are associated with the parameters $\xi\in\{\xi_{\gamma_+}\}$. The remaining lines and columns are associated with $\xi\in\{\xi_{\gamma_-}\}$.*

Thus, we have obtained an explicit answer for the generating function $\langle Q_\kappa(m)\rangle$ of the inhomogeneous $XXZ$ model. It is also possible to check that the above sum over partitions remains indeed finite in the homogeneous limit $\xi_k\to 0$.

**5.2. Asymptotic results.** An important issue is the analysis of the multiple integral representations of correlation functions for large distances. There it means analyzing asymptotic behavior of $m$-fold integrals for $m$ large. An interesting example to study in this respect is provided by the emptiness formation probability. This correlation function reduces to a single elementary block. Moreover, we already described its exact value for an anisotropy $\Delta=\frac{1}{2}$ in the previous section. In fact, it is possible to obtain the asymptotic behavior of $\tau(m)$ using the saddle-point method for arbitrary values of the anisotropy $\Delta>-1$ . This was performed for the first time in [12] in the case of free fermions ($\Delta=0$), but it can be applied to the general case as well. We present here the results in the massless and massive regimes [14], [19].

To apply the saddle-point method to the emptiness formation probability, it is convenient to express its integral representation in the following form:

$$\tau(m) = \int_{\mathcal{D}} d^m\lambda\; G_m(\{\lambda\})\; e^{m^2 S_m(\{\lambda\})}, \tag{67}$$

with

$$\begin{aligned} S_m(\{\lambda\}) = &-\frac{1}{m^2}\sum_{a>b}^{m}\log[\sinh(\lambda_a-\lambda_b+\eta)\ \sinh(\lambda_a-\lambda_b-\eta)] \\ &+\frac{1}{m}\sum_{a=1}^{m}\log[\sinh(\lambda_a+\eta/2)\ \sinh(\lambda_a-\eta/2)] \\ &+\frac{1}{m^2}\lim_{\xi_1\ldots\xi_m\to\eta/2}\log\Big[\Big(\frac{-2i\pi}{\sinh\eta}\Big)^m\frac{\big(\det\rho(\lambda_j,\xi_k)\big)^2}{\prod\limits_{a\neq b}\sinh(\xi_a-\xi_b)}\Big] \end{aligned} \tag{68}$$

and

$$G_m(\{\lambda\}) = \lim_{\xi_1\ldots\xi_m\to\eta/2}\frac{\det_m\left[\frac{i}{2\pi}t(\lambda_j,\xi_k)\right]}{\det_m\rho(\lambda_j,\xi_k)}. \tag{69}$$

In (67), the integration domain $\mathcal{D}$ is such that the variables of integration $\lambda_1,\ldots,\lambda_m$ are ordered in the interval $\mathcal{C}=[-\Lambda_h,\Lambda_h]$ (i.e. $-\Lambda_h<\lambda_1<\cdots<\lambda_m<\Lambda_h$ in the massless case, and $-i\Lambda_h<i\lambda_1<\cdots<i\lambda_m<i\Lambda_h$ in the massive case).

The main problem in the saddle point analysis is that, a priori, we do not know any asymptotic equivalent of the quantity $G_m(\lambda)$ when $m\to\infty$. Nevertheless, in the case of zero magnetic field, it is still possible to compute the asymptotic behavior of (67) in the leading order, provided we make the following hypothesis: we assume that the integrand of (67) admits a maximum for a certain value $\lambda'_1,\ldots,\lambda'_m$ of the integration variables $\lambda_1,\ldots,\lambda_m$ such that, for large $m$, the distribution of these parameters $\lambda'_1,\ldots,\lambda'_m$ can be described by a density function $\rho_s(\lambda')$ of the form

$$\rho_s(\lambda'_j)=\lim_{m\to\infty}\frac{1}{m(\lambda'_{j+1}-\lambda'_j)} \tag{70}$$

on the symmetric interval $[-\Lambda,\Lambda]$, and that, at the leading order in $m$, we can replace the sums over the set of parameters $\{\lambda'\}$ by integrals weighted with the density $\rho_s(\lambda')$.

First, it is easy to determine the maximum of the function $S_m(\{\lambda\})$. Indeed, let $\{\tilde{\lambda}\}$ be solution of the system

$$\partial_{\lambda_j}S_m(\{\tilde{\lambda}\})=0,\quad 1\le j\le m. \tag{71}$$

In the limit $m\to\infty$, if we suppose again that the parameters $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_m$ become distributed according to a certain density $\tilde{\rho}_s(\lambda)$ and that sums over the $\tilde{\lambda}_j$ become integrals over this density, the system (71) turns again into a single integral equation for $\tilde{\rho}_s$, that can be solved explicitly in the case of zero magnetic field. It gives the maximum of $S_m(\{\lambda\})$ when $m\to\infty$[2].

[2]At this main order in $m$, there exists a unique solution of the integral equation for $\tilde{\rho}_s$, and we know it corresponds to a maximum because $S_m(\{\lambda\})\to-\infty$ on the boundary of $\mathcal{D}$.

The second step is to show that the factor $G_m(\{\lambda\})$ gives always a negligible contribution compared to $S_m(\{\tilde{\lambda}\})$ at this order in $m$, at least for any distribution of the variables $\lambda_j$ satisfying the previous hypothesis of regularity. We obtain

$$\lim_{m\to\infty} \frac{1}{m^2} \log G_m(\{\lambda\}) = 0 \tag{72}$$

for any distribution of $\{\lambda\}$ with good properties of regularity, in particular for the saddle point. This means that, at the main order in $m$, the factor $G_m(\{\lambda\})$ does not contribute to the value of the maximum of the integrand.

Finally we obtain the following result concerning the asymptotic behaviour of $\tau(m)$ for $m \to \infty$ (see [14], [19]):

$$S^{(0)}(\Delta) = \lim_{m\to\infty} \frac{\log \tau(m)}{m^2}, \tag{73}$$

$$= -\frac{\zeta}{2} - \sum_{n=1}^{\infty} \frac{e^{-n\zeta}}{n} \frac{\sinh(n\zeta)}{\cosh(2n\zeta)}, \qquad (\Delta = \cosh\zeta > 1), \tag{74}$$

$$= \log\frac{\pi}{\zeta} + \frac{1}{2} \int_{\mathbb{R}-i0} \frac{d\omega}{\omega} \frac{\sinh\frac{\omega}{2}(\pi-\zeta)\cosh^2\frac{\omega\zeta}{2}}{\sinh\frac{\pi\omega}{2}\sinh\frac{\omega\zeta}{2}\cosh\omega\zeta}, \quad (-1 < \Delta = \cos\zeta < 1). \tag{75}$$

It coincides with the exact known results obtained in [46], [12] at the free fermion point and in [45], [13] at $\Delta = 1/2$, and is in agreement with the expected (infinite) value in the Ising limit. Similar techniques can be applied to the two point function. However, the result that has been extracted so far is only the absence of the gaussian term. Unfortunately, we do not know up to now how to extract the expected power law corrections to the gaussian behavior from this saddle point analysis. More powerful methods will certainly be needed to go further.

## Conclusion and perspectives

In this article, we have reviewed recent results concerning the computation of correlation functions in the $XXZ$ chain by the methods of the inverse scattering problem and the algebraic Bethe ansatz. In conclusion, we would like to discuss some perspectives and problems to be solved.

One of the most interesting open problems is to prove the conformal field theory predictions [47], [48], [49], [50] concerning the asymptotic behavior of the correlation functions. This is certainly a very important issue not only for physical applications but also from a theoretical view point. Moreover, it also would open the route towards the generalization of the methods presented here to quantum integrable models of field theory. We have seen that in particular cases, the multiple integral representations

enable for a preliminary asymptotic analysis. Nevertheless, this problem remains one of the main challenges in the topics that have been described in this article.

A possible way to solve this problem would be to find the thermodynamic limit of the master equations (like the one obtained for the two point correlation functions). It is natural to expect that, in this limit, one should obtain a representation for these correlation functions in terms of a functional integral, which could eventually be estimated for large time and distance.

Note that the master equation shows a direct analytic relation between the multiple integral representations and the form factor expansions for the correlation functions. It seems likely that similar representations exist for other models solvable by algebraic Bethe ansatz. It would be in particular very interesting to obtain an analogue of this master equation in the case of the field theory models, which could provide an analytic link between short distance and long distance expansions of their correlation functions.

## References

[1] Baxter, R. J., *Exactly solved models in statistical mechanics*. Academic Press, London, New York 1982.

[2] Gaudin, M., *La fonction d'onde de Bethe*. Masson, Paris 1983.

[3] Lieb, E. H., and Mattis, D. C., *Mathematical Physics in One Dimension*. Academic Press, New York 1966.

[4] Thacker, H. B., Exact integrability in quantum field theory and statistical systems. *Rev. Mod. Phys.* **53** (1981), 253.

[5] Korepin, V. E., Bogoliubov, N. M., and Izergin, A. G., *Quantum inverse scattering method and correlation functions*. Cambridge University Press, Cambridge 1993.

[6] Jimbo, M., and Miwa, T., *Algebraic analysis of solvable lattice models*. CBMS Reg. Conf. Ser. Math. 85, Amer. Math. Soc., Providence, RI, 1995.

[7] Maillet, J. M., and Sanchez de Santos, J., Drinfeld twists and algebraic Bethe ansatz. *Amer. Math. Soc. Transl.* **201** (2000), 137.

[8] Kitanine, N., Maillet, J. M., and Terras, V., Form factors of the $XXZ$ Heisenberg spin-1/2 finite chain. *Nucl. Phys. B* **554** (1999), 647.

[9] Kitanine, N., Maillet, J. M., and Terras, V., Correlation functions of the $XXZ$ Heisenberg spin-1/2 chain in a magnetic field. *Nucl. Phys. B* **567** (2000), 554.

[10] Maillet, J. M., and Terras, V., On the quantum inverse scattering problem. *Nucl. Phys. B* **575** (2000), 627.

[11] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Spin–spin correlation functions of the $XXZ$-1/2 Heisenberg chain in a magnetic field. *Nucl. Phys. B* **641** (2002), 487.

[12] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Correlation functions of the $XXZ$ spin-1/2 Heisenberg chain at the free fermion point from their multiple integral representations. *Nucl. Phys. B* **642** (2002), 433.

[13] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Emptiness formation probability of the $XXZ$ spin-1/2 Heisenberg chain at Delta $= 1/2$. *J. Phys. A* **35** (2002), L385.

[14] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Large distance asymptotic behaviour of the emptiness formation probability of the $XXZ$ spin-1/2 Heisenberg chain. *J. Phys. A* **35** (2002), L753.

[15] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Master equation for spin-spin correlation functions of the $XXZ$ chain. *Nucl. Phys. B* **712** [FS] (2005), 600.

[16] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Dynamical correlation functions of the $XXZ$ spin-1/2 chain. *Nucl. Phys. B* **729** [FS] (2005), 558.

[17] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., On the spin-spin correlation functions of the $XXZ$ spin-1/2 infinite chain. *J.Phys. A* **38** (2005), 7441.

[18] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., Exact results for the $\sigma^z$ two-point function of the $XXZ$ chain at $\Delta = 1/2$. *J. Stat. Mech.* (2005), L09002 (electronic).

[19] Kitanine, N., Maillet, J. M., Slavnov, N. A., and V. Terras, V., On the algebraic Bethe Ansatz approach to the correlation functions of the $XXZ$ spin-1/2 Heisenberg chain. In *Recent Progress in Solvable Lattice Models*, RIMS Sciences Project Research 2004 on Method of Algebraic Analysis in Integrable Systems, RIMS, Kyoto 2004.

[20] Heisenberg, H., Zur Theorie der Ferromagnetismus. *Z. Phys.* **49** (1928), 619.

[21] Bethe, H., Zur Theorie der Metalle I. Eigenwerte und Eigenfunktionen Atomkete. *Z. Phys.* **71** (1931), 205.

[22] Orbach, R., Linear antiferromagnetic chain with anisotropic coupling. *Phys. Rev.* **112** (1958), 309.

[23] Walker, L. R., Antiferromagnetic linear chain. *Phys. Rev.* **116** (1959), 1089.

[24] Yang, C. N. and Yang, C. P., One-dimensional chain of anisotropic spin-spin interactions. I. Proof of Bethe's hypothesis for ground state in a finite system. *Phys. Rev.* **150** (1966), 321.

[25] Yang, C. N. and Yang, C. P., One-dimensional chain of anisotropic spin-spin interactions. II. Properties of the ground state energy per lattice site for an infinite system. *Phys. Rev.* **150** (1966), 327.

[26] Faddeev, L. D., Sklyanin, E. K., and Takhtajan, L. A., Quantum inverse problem method I. *Theor. Math. Phys.* **40** (1980), 688.

[27] Takhtajan, L. A., and Faddeev, L. D., The Quantum method of the inverse problem and the Heisenberg $XYZ$ model. *Russ. Math. Surveys* **34** (1979), 11.

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

[29] Lieb, E., Schultz, T., and Mattis, D., Two soluble models of an antiferromagnetic chain. *Ann. Phys.* **16** (1961), 407.

[30] McCoy, B., Spin correlation functions of the $XY$ model. *Phys. Rev.* **173** (1968), 531.

[31] Wu, T. T., McCoy, B. M., Tracy, C. A., and Barouch, E., The spin-spin correlation function of the two-dimensional Ising model: exact results in the scaling region. *Phys. Rev. B* **13** (1976), 316.

[32] McCoy, B. M., Tracy, C. A., and Wu, T. T., Two-dimensional Ising model as an exactly soluble relativistic quantum field theory: explicit formulas for the $n$-point functions. *Phys. Rev. Lett.* **38** (1977), 793.

[33] Sato, M., Miwa, T., and Jimbo, M., Holonomic quantum fields. *Publ. Res. Inst. Math. Sci.* **14** (1978), 223; **15** (1979), 201, 577, 871; **16** (1980), 531.

[34] Jimbo, M., Miki, K., Miwa, T., and Nakayashiki, A., Correlation functions of the $XXZ$ model for $\Delta < -1$. *Phys. Lett. A* **168** (1992), 256.

[35] Jimbo, M., and Miwa, T., Quantum KZ equation with $|q| = 1$ and correlation functions of the $XXZ$ model in the gapless regime. *J. Phys. A.* **29** (1996), 2923.

[36] Göhmann, F., Klümper, A., and Seel, A., Integral representations for correlation functions of the $XXZ$ chain at finite temperature. *J. Phys. A* **37** (2004), 7625.

[37] Boos, H., Jimbo, M., Miwa, T., Smirnov, F., and Takeyama, Y., A recursion formula for the correlation functions of an inhomogeneous $XXX$ model. hep-th/0405044, 2004.

[38] Boos, H., Jimbo, M., Miwa, T., Smirnov, F., and Takeyama, Y., Reduced q-KZ equation and correlation functions of the $XXZ$ model. hep-th/0412191, 2004.

[39] Tarasov V., and Varchenko, A., Completeness of Bethe vectors and difference equations with regular singular points. *Internat. Math. Res. Notices* **13** (1996), 637.

[40] Drinfel'd, V. G., Quantum Groups. In *Proceedings of the International Congress of Mathematicians* (Berkeley, Calif., 1986), Vol. 1, Amer. Math. Soc., Providence, R.I., 1987, 798–820.

[41] Izergin, A. G., Partition function of the six-vertex model in a finite volume. *Sov. Phys. Dokl.* **32** (1987), 878.

[42] Slavnov, N. A., Calculation of scalar products of wave functions and form factors in the framework of the algebraic Bethe Ansatz. *Theor. Math. Phys.* **79** (1989), 502.

[43] Gaudin, M., Mc Coy, B. M., and Wu, T. T., Normalization sum for the Bethe's hypothesis wave functions of the Heisenberg-Ising model. *Phys. Rev. D* **23** (1981), 417.

[44] Korepin, V. E., Calculation of norms of Bethe wave functions. *Commun. Math. Phys.* **86** (1982), 391.

[45] Razumov, A. V. and Stroganov, Y. G., Spin chains and combinatorics. *J. Phys. A* **34** (2001), 3185.

[46] M. Shiroishi, M. Takahashi, and Y. Nishiyama, Emptiness Formation Probability for the One-Dimensional Isotropic $XY$ Model. *J. Phys. Soc. Jap.* **70** (2001), 3535.

[47] Luther, A., and Peschel, I., Calculation of critical exponents in two dimensions from quantum field theory in one dimension. *Phys. Rev. B* **12** (1975), 3908.

[48] Belavin, A. A., Polyakov, A. M., and Zamolodchikov, A. B., Infinite conformal symmetry in two-dimensional quantum field theory. *Nucl. Phys. B* **241** (1984), 333.

[49] Lukyanov, S., Correlation amplitude for the $XXZ$ spin chain in the disordered regime. *Phys. Rev. B* **59** (1999), 11163.

[50] Lukyanov, S., Terras, V., Long-distance asymptotics of spin-spin correlation functions for the $XXZ$ spin chain. *Nuclear Phys. B* **654** (2003), 323.

Laboratoire de Physique, ENS Lyon, UMR 5672 du CNRS, 46 Allée d'Italie, 69364 Lyon, France

E-mail: maillet@ens-lyon.fr

# Gromov–Witten invariants and topological strings: a progress report

Marcos Mariño*

**Abstract.** In this talk I summarize recent progress in the theory of Gromov–Witten invariants from topological string theory and string dualities. On the one hand, large $N$ dualities have led to the theory of the topological vertex, which solves Gromov–Witten theory to all genera on toric, noncompact Calabi–Yau threefolds. On the other hand, heterotic/type II duality and the holomorphic anomaly equations can be used to analyze Gromov–Witten theory in some simple compact examples. I sketch the physical ideas behind these results and connect the results obtained in physics with the results obtained in algebraic geometry.

**Mathematics Subject Classification (2000).** Primary 00A05; Secondary 00B10.

**Keywords.** String theory, Gromov–Witten invariants.

## 1. Introduction

The theory of Gromov–Witten invariants was largely motivated by the study of string theory on Calabi–Yau manifolds, and has now developed into one of the most dynamic fields of algebraic geometry. During the last years there has been enormous progress in the development of the theory and of its computational techniques. Roughly speaking, and restricting ourselves to Calabi–Yau threefolds, we have the following mathematical approaches to the computation of Gromov–Witten invariants:

1. *Localization*. This was first proposed by Kontsevich, and requires torus actions in the Calabi–Yau in order to work. Localization provides *a priori* a complete solution of the theory on toric (hence non-compact) Calabi–Yau manifolds, and reduces the computation of Gromov–Witten invariants to the calculation of Hodge integrals in Deligne–Mumford moduli space. Localization techniques make also possible to solve the theory at genus zero on a wide class of compact manifolds, see [8], [14] for a review.

2. *Deformation and topological approach*. This has been developed more recently and relies on *relative* Gromov–Witten invariants [11], [24]. It provides a cut-and-paste approach to the calculation of the invariants and seems to be the most

*I am grateful to M. Aganagic, A. Klemm and C. Vafa for collaboration on the work reported here. I would also like to thank D. Maulik, G. Moore and R. Pandharipande for discussions.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

powerful approach to higher genus Gromov–Witten invariants in the compact case.

3. *D-brane moduli spaces*. Gromov–Witten invariants can be reformulated in terms of the so-called Gopakumar–Vafa invariants (see [14], [21] for a summary of these). Heuristic techniques to compute these have been developed in [16], as Euler characteristics of moduli space of embedded surfaces, and one can recover to a large extent the original information of Gromov–Witten theory. The equivalence between these two invariants remains however conjectural, and a general, rigorous definition of the Gopakumar–Vafa invariants in terms of appropriate moduli spaces is still not known.

4. *Equivalence to Donaldson–Thomas invariants*. In [23] it was proposed that Gromov–Witten invariants are equivalent to Donaldson–Thomas invariants, which are associated to moduli spaces of sheaves. This equivalence remains largely conjectural and so far it has led to little computational progress, although it is currently an area of active research.

Gromov–Witten invariants are closely related to string theory. It turns out that type IIA theory on a Calabi–Yau manifold $X$ leads to a four-dimensional supersymmetric theory whose Lagrangian contains moduli-dependent couplings $F_g(t)$, where $t$ denotes the Kähler moduli of the Calabi–Yau. When these couplings are expanded in the large radius limit, they are of the form

$$F_g(t) = \sum_{Q \in H_2(X)} N_{g,Q}\, \mathrm{e}^{-Q\cdot t}, \tag{1}$$

where $N_{g,Q}$ are the Gromov–Witten invariants for the class $Q$ at genus $g$. It turns out that there is a simplified version of string theory, called topological string theory, which captures precisely the information contained in these couplings. Topological string theory comes in two versions, called the A and the B model (see [21], [14] for a review). Type A topological string theory is related to Gromov–Witten theory, and its free energy at genus $g$ is precisely given by (1). Type B topological string theory is related to the deformation theory of complex structures of the Calabi–Yau manifold. In the last years, various dualities of string theory have led to powerful techniques to compute these couplings, hence Gromov–Witten invariants:

1. *Mirror symmetry*. Mirror symmetry relates type A theory on a Calabi–Yau manifold $X$ to type B theory on the mirror manifold $\widetilde{X}$. When the mirror of the Calabi–Yau $X$ is known, this leads to a complete solution at genus zero in terms of variation of the complex structures of $\widetilde{X}$. For genus $g \geq 1$, mirror symmetry can be combined with the holomorphic anomaly equations of [5] to obtain $F_g(t)$. However, this does not provide the full solution to the model due to the so-called holomorphic ambiguity. On the other hand, mirror symmetry and the holomorphic anomaly equation are very general and work for both compact and non-compact Calabi–Yau manifolds.

2. *Large N dualities*. Large $N$ dualities lead to a computation of the $F_g(t)$ couplings in terms of correlation functions and partition functions in Chern–Simons theory. Although this was formulated originally only for the resolved conifold, one ends up with a general theory – the theory of the topological vertex, introduced in [2] – which leads to a complete solution on toric Calabi–Yau manifolds. The theory of the topological vertex is closely related to localization and to Hodge integrals, and it can be formulated in a rigorous mathematical way [19], [23].

3. *Heterotic duality*. When the Calabi–Yau manifold has the structure of a K3 fibration, type IIA theory often has a heterotic dual, and the evaluation of $F_g(t)$ restricted to the K3 fiber can be reduced to a one-loop integral in heterotic string theory [4], [22]. This leads to explicit, conjectural formulae for Gromov–Witten invariants in terms of modular forms.

In this progress report I will concentrate on two results: (1) I will summarize how large $N$ dualities lead to a complete solution of topological string theory on toric Calabi–Yau manifolds. (2) I will discuss what is probably the simplest, non-trivial compact Calabi–Yau manifold, the so-called Enriques Calabi–Yau manifold, which is very tractable both mathematically and physically, and might be the natural starting point to understand the compact case.

## 2. The toric case

In this section we give a rather brief summary of the results obtained in the context of topological string theory to compute Gromov–Witten invariants of toric geometries. This subject has been extensively reviewed in [20], [21], to which we refer for further information and/or references.

**2.1. The Gopakumar–Vafa duality.** The Gopakumar–Vafa duality [12] is an example of the string/gauge theory dualities which have been discovered in the last years. It relates a particular gauge theory – $U(N)$ Chern–Simons theory on the three-sphere with coupling $k$ – to a particular string theory – the type A topological string on the small resolution of the conifold singularity. This is a toric (hence non-compact) Calabi–Yau manifold which can be regarded as the total space of the bundle

$$\mathcal{O}(-1) \oplus \mathcal{O}(-1) \to \mathbb{P}^1. \tag{2}$$

It has a single Kähler parameter $t$ which gives the complexified area of the $\mathbb{P}^1$. The identification is such that the effective coupling constant of the gauge theory

$$g = \frac{2\pi \mathrm{i}}{k+N} \tag{3}$$

is identified to the string coupling constant $g_s$, while the 't Hooft parameter $gN$ is identified with the Kähler parameter $t$:

$$t = \frac{2\pi \mathrm{i} N}{k+N}. \tag{4}$$

In particular, the duality asserts that the free energy of Chern–Simons theory on $\mathbb{S}^3$ equals the total free energy of the topological string, which is defined by summing the topological string amplitudes to all genera,

$$F = \sum_{g=0}^{\infty} F_g(t) g_s^{2g-2}, \tag{5}$$

This can be checked explicitly since both quantities are known. The free energies $F_g(t)$ of topological string theory on the resolved conifold have been computed in various ways (see for example [7]), and the free energy of Chern–Simons theory on the sphere was computed by Witten and is given by

$$F = \log Z = \log\ S_{00}, \tag{6}$$

where

$$S_{00} = \prod_{j=1}^{N} \left( 2 \sin \frac{2\pi \mathrm{i} j}{k+N} \right)^{N-j} \tag{7}$$

is obtained from the theory of affine Lie algebras.

The duality of Gopakumar and Vafa gives some important information on Gromov–Witten theory, but it only deals with one particular Calabi–Yau manifold: the resolved conifold. From the point of view of topological string theory it would be extremely useful to have generalizations of the duality which cover other situations, and express the amplitude $F_g(t)$ in terms of gauge theoretic quantities. It turns out that this can be done in two different ways, which I consider in the next two subsections.

**2.2. Extensions to other geometries.** The first possibility to generalize the Gopakumar–Vafa duality consists on taking Chern–Simons theory on more general three-manifolds and search for Calabi–Yau duals. One obvious way to achieve this is to do a quotient of both sides of the duality by $\mathbb{Z}_p$ symmetry. On the gauge theory side one obtains Chern–Simons theory on the lens space $L(p, 1)$. The quotient of the resolved conifold leads to a toric geometry, the $A_{p-1}$ fibration over $\mathbb{P}^1$, which has $p$ Kähler parameters. For example, for $p = 2$ this leads to a duality between Chern–Simons theory on $\mathbb{RP}^3$ and topological string theory on the Calabi–Yau manifold given by the anticanonical bundle of $\mathbb{P}^1 \times \mathbb{P}^1$.

This generalization of the Gopakumar–Vafa duality was proposed in [1], where it was tested in detail for $p = 2$. One interesting aspect of it is that one has to consider

the Chern–Simons theory around an arbitrary reducible flat connection which breaks the gauge group

$$U(N) \to \prod_{i=1}^{p} U(N_i). \tag{8}$$

The $p$ Kähler parameters of the Calabi–Yau manifold, $t_i$, are identified with the partial 't Hooft parameters of the Chern–Simons theory with broken gauge symmetry:

$$t_i = g_s N_i, \quad i = 1, \cdots, p. \tag{9}$$

Unfortunately, it is not known if there are further generalizations of these results. However, it seems natural to state the following

**Conjecture 2.1.** Given a rational homology sphere $M$, there exists a Calabi–Yau manifold $X_M$ such that the free energy of Chern–Simons theory on $M$, expanded around a generic reducible flat connection, equals the total free energy of topological string theory on $X_M$.

**2.3. The cut-and-paste approach: the topological vertex.** The basic idea of the topological vertex is to regard a generic toric geometry as made out of pieces where one can use the duality of Gopakumar and Vafa between the resolved conifold and Chern–Simons theory. A first approach is then to cut the manifold into pieces that are locally like resolved conifolds, to associate a topological string amplitude to each of the pieces, and then to glue the results together. This program was developed in [3], [9]. It turns out that there is a natural way to cut the original manifold into pieces, and this is by introducing D-branes around Lagrangian submanifolds. The amplitude associated to each of the pieces is then, due to the presence of D-branes, an *open* topological string amplitude. Fortunately, the duality of Gopakumar and Vafa also holds in the open setting [26], and the open amplitudes are closely related to *Chern–Simons invariants of knots and links* in $\mathbb{S}^3$. One then finds a surprising relation between these invariants and Gromov–Witten invariants of toric Calabi–Yau threefolds.

This procedure was refined and generalized in [2]. In the approach of [3], [9], one has to divide the geometry into pieces which are like the resolved conifold, which in terms of toric diagrams can be regarded as a four-valent graph. However, the basic building block is in fact a *trivalent* vertex that corresponds to $\mathbb{C}^3$ with three sets of D-branes wrapping Lagrangian submanifolds. The open topological string amplitude associated to this graph is called the topological vertex, and it is denoted by

$$C_{R_1 R_2 R_3}, \tag{10}$$

where $R_i$ are representations of $U(\infty)$ (or, equivalently, Young tableaux) which correspond roughly to the Chan–Paton factors associated to the open strings ending on the branes. The topological vertex depends only on the string coupling $g_s$, and from

its power series expansion in $g_s$ one can extract open Gromov–Witten invariants of $\mathbb{C}^3$ with specified Lagrangian boundary conditions. These invariants do not have a rigorous mathematical definition, but they can be re-interpreted as *relative* Gromov–Witten invariants and computed by localization [19]. As shown in [2], one can use a subtle variant of the Gopakumar–Vafa duality to obtain an explicit expression for the topological vertex in terms of known quantum group invariants for arbitrary representations $R_i$. The final result is

$$C_{R_1R_2R_3} = q^{\frac{\kappa_{R_2}+\kappa_{R_3}}{2}} \sum_{Q_1,Q_3,Q} N_{QQ_1}^{R_1} N_{QQ_3}^{R_3^t} \frac{W_{R_2^tQ_1} W_{R_2Q_3}}{W_{R_2}}. \tag{11}$$

In this equation, $N_{R_1R_2}^R$ are Littlewood–Richardson coefficients, and $R^t$ denotes the transpose of the representation $R$. $\kappa_R$ is related to the second Casimir of $R$ as a representation of $U(N)$ and can be written as

$$\kappa_R = \ell(R) + \sum_i l_i^R (l_i^R - 2i), \tag{12}$$

where $\ell(R)$ is the number of boxes in the Young tableau of $R$ and $l_i^R$ is the number of boxes in the $i$-th row of $R$. Finally, $W_{R_1R_2}$ are related to quantum group invariants of the Hopf link and can be written in terms of Schur polynomials as

$$W_{R_1R_2}(q) = s_{R_2}\big(x_i = q^{-i+\frac{1}{2}}\big) s_{R_1}\big(x_i = q^{l_i^{R_2}-i+\frac{1}{2}}\big), \tag{13}$$

where $q = \mathrm{e}^{g_s}$.

In [2] it is shown on a physical basis that the all genus Gromov–Witten invariants of any toric Calabi–Yau manifold can be computed from (11) and some simple gluing rules. Essentially, one takes the toric diagram of the Calabi–Yau and decomposes it into trivalent vertices. To each of these vertices one assigns the amplitude (11), and then one glues them together according to simple instructions encoded in the diagram. The result is the partition function $Z = \exp F$, where $F$ is the free energy (5). One simple example is the so-called local $\mathbb{P}^2$ manifold, namely the total space of the bundle $\mathcal{O}(-3) \to \mathbb{P}^2$. The rules of [2] give

$$Z_{\mathbb{P}^2} = \sum_{R_1,R_2,R_3} (-1)^{\sum_i \ell(R_i)} e^{-\sum_i \ell(R_i)t} q^{-\sum_i \kappa_{R_i}} C_{0R_2^tR_3} C_{0R_1^tR_2} C_{0R_3^tR_1}. \tag{14}$$

The theory of the topological vertex developed in [2] and largely based on the ideas of string/gauge theory duality has been re-derived to a large extent on a rigorous mathematical basis in [19], by using relative Gromov–Witten invariants. This theory gives a full solution of topological string theory in the toric case. Let us now consider the compact case.

## 3. The compact case

**3.1. Heterotic duality.** As we already explained in the introduction, the main tool to compute Gromov–Witten invariants in the compact case is the combination of mirror symmetry and the holomorphic anomaly equation of [5]. This approach is very general, but it does not give a complete solution to the problem of computing the topological string amplitudes due to the holomorphic ambiguity. It turns out that if the Calabi–Yau manifold has the structure of a K3 fibration over $\mathbb{P}^1$, one can do better. The reason is that type IIA theory on such manifolds has a heterotic dual [15], and for Kähler classes in the K3 fiber, one can compute the $F_g(t)$ couplings by doing a one-loop computation in the heterotic string. This leads to close expressions for the topological string amplitudes in terms of modular forms [13], [22].

Let us consider for example the STU model first studied in [15]. This is a K3 fibration where the K3 fiber has two complexified Kähler parameters, $t = (t^+, t^-)$ with $\operatorname{Re} t^\pm > 0$. The Kähler classes are labelled by two integers $r = (n, m)$. These classes form a lattice $\Gamma^{1,1}$ with intersection form

$$H = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{15}$$

which defines an inner product such that $r \cdot t = mt^+ + nt^-$ and $r^2 = 2nm$. An involved heterotic computation [22] leads to a topological string coupling $F_g(t)$ of the form

$$F_g(t) = \sum_{r>0} c_g(r^2/2)\mathrm{Li}_{3-2g}(\mathrm{e}^{-r\cdot t}). \tag{16}$$

In this formula, $r > 0$ means the following possibilities: $(n, m) = (1, -1)$, $n > 0$, $m > 0$, $n = 0$, $m > 0$, or $n > 0$. The coefficients $c(n)$ are defined by

$$\sum_n c_g(n)q^n = -\frac{2E_4(q)E_6(q)}{\eta^{24}(q)}\mathcal{P}_g(q), \tag{17}$$

where $E_4$, $E_6$ are Eisenstein series, $\eta$ is the Dedekind eta function, and the polynomials $\mathcal{P}_g(q)$ are given by

$$\left(\frac{2\pi\eta^3\lambda}{\vartheta_1(\lambda|\tau)}\right)^2 = \sum_{g=0}^{\infty}(2\pi\lambda)^{2g}\mathcal{P}_g(q), \tag{18}$$

and can be written as polynomials in the Eisenstein series $E_2$, $E_4$ and $E_6$ (see [22], [18] for more details on the modular forms involved).

Other examples of heterotic computations of topological string amplitudes can be found in [17], which also tested the predictions by using the holomorphic anomaly equations and extended them up to genus two by including the "missing" Kähler parameter on the base of the fibration.

**3.2. A very special example: the Enriques Calabi–Yau.** Although many techniques have been developed in order to compute topological string amplitudes in the compact case, typically the theory becomes intractable at high genus and/or degree. In this sense, it would be important to identify *the* compact CY manifold where topological string theory is most tractable.

There is a compact example where topological string theory is exactly solvable, namely $\mathrm{K3}\times\mathbb{T}^2$. The topological string amplitudes are however rather trivial in this case, and in particular they vanish for $g \geq 2$. Hence this example is too simple, and this is due to the extended $\mathcal{N} = 4$ supersymmetry of the corresponding type II theory, related in turn to the SU(2) holonomy. It is then natural to consider Calabi–Yau manifolds with holonomy $H$ which is intermediate between the SU(2) and the generic one SU(3). A Calabi–Yau manifold with intermediate holonomy $\mathrm{SU}(2)\times\mathbb{Z}_2$ has been constructed in [6], [27], [10] as an orbifold w.r.t. a free $\mathbb{Z}_2$ involution of $\mathrm{K3}\times\mathbb{T}^2$. The resulting space exhibits a K3 fibration with four fibers of multiplicity two over the four fixed points of the involution in the base. These fibers are Enriques surfaces. A good deal of the nontrivial geometry of this CY comes from the geometry of the Enriques fibers, and therefore this example has been called the *Enriques CY manifold*. The string model obtained by compactifying type II theory on the Enriques CY has $\mathcal{N} = 2$ supersymmetry and is known as the FHSV model. The Enriques CY seems to be the simplest CY compactification with nontrivial topological string amplitudes. Moreover it has a dual description as an asymmetric orbifold of the heterotic string [10].

A first step in order to determine the topological string amplitudes is then to compute the amplitudes on the fiber, by using the heterotic dual and techniques similar to those of [22]. In order to write down the result, we notice that the Enriques fiber has ten Kähler parameters $t = (t^+, t^-, \vec{t}\,)$. The Kähler classes belong to the cone $\Gamma_E = \Gamma^{1,1} \oplus E_8(-1)$, and will be parametrized by a vector of integer numbers $r = (n, m, \vec{q}\,)$. The topological string amplitudes on the Kähler cone are given by

$$F_g^E(t) = \sum_{r>0} c_g(r^2)\left\{2^{3-2g}\mathrm{Li}_{3-2g}(e^{-r\cdot t}) - \mathrm{Li}_{3-2g}(e^{-2r\cdot t})\right\}, \tag{19}$$

where

$$\sum_n c_g(n)q^n = -\frac{2}{q}\prod_{n=1}^{\infty}(1-q^{2n})^{-12}\mathcal{P}_g(q), \tag{20}$$

and $r^2 = 2mn - \vec{q}^{\,2}$. The restriction $r > 0$ means now that $n > 0$, or $n = 0, m > 0$, or $n = m = 0, \vec{q} > 0$. The superscript $E$ refers to the Enriques fiber.

The Enriques Calabi–Yau manifold has an extra Kähler class corresponding to the $\mathbb{P}^1$ in the base of the fibration. The perturbative heterotic string theory does not give information on the dependence of $F_g(t)$ on this parameter, and the only available technique to do that for the moment being is the holomorphic anomaly equation. It turns out, however, that the Enriques Calabi–Yau is particularly simple in that respect, and one can solve the holomorphic anomaly equation at low genera. In this way one

finds explicit expressions for $F_1$ and $F_2$ on the *total* Calabi–Yau. If we denote by $S$ the Kähler parameter of the base, one finds

$$\begin{aligned} F_1(t, S) &= F_1^E(t) - 12 \log\, \eta(q_S), \\ F_2(t, S) &= E_2(q_S) F_2^E(t), \end{aligned} \tag{21}$$

where $q_S = \mathrm{e}^{-S}$. Notice that, as $S \to \infty$, one recovers the results in the fiber. These results are surprisingly simple, and they have been verified in [25] by using the topological techniques in Gromov–Witten theory. This is the only compact manifold where the topological string amplitudes have been computed up to genus two *both* in the context of topological string theory and in the context of algebraic geometry. It seems to be the most accessible compact example and it provides a fruitful interaction between physical and mathematical techniques. Results for genus 3 and 4 including the Kähler parameter of the base have also been obtained in [18]. One important remaining question is: is the Enriques Calabi–Yau an exactly solvable example?

## 4. Conclusions

One obvious and general conclusion is that the interaction between topological string theory and Gromov–Witten theory and algebraic geometry has been extremely rich and rewarding for both fields. String dualities have led to the solution of many models and to sometimes surprising mathematical predictions, while rigorous mathematical techniques have confirmed many of the physical ideas underlying the predictions. This interplay has led so far to a rather complete solution of the problem in the toric case. Progress in the compact case has been also significant, although the problem seems to be much harder and we are lacking an effective strategy to address general models. For this reason, we have proposed in [18] to focus on relatively simple examples where the theory might have some simplifying features yet be rich enough to display the essential complexities of the problem. So far, the simplest nontrivial valley in the landscape of compact Calabi–Yau's seems to be the Enriques Calabi–Yau, which has been solved at low genera in [18], [25]. The main open problem is of course to find an organizing principle that makes possible to address the general compact case in an effective way.

## References

[1] Aganagic, M., Klemm, A., Mariño, M., Vafa, C., Matrix model as a mirror of Chern–Simons theory. *J. High Energy Phys.* **0402** (2004), 010.

[2] Aganagic, M., Klemm, A., Mariño, M., Vafa, C., The topological vertex. *Comm. Math. Phys.* **254** (2005), 425–478.

[3] Aganagic, M., Mariño, M., Vafa, C., All loop topological string amplitudes from Chern–Simons theory. *Comm. Math. Phys.* **247** (2004), 467–512.

[4] Antoniadis, I., Gava, E., Narain, K. S., Taylor, T. R., $N = 2$ type II heterotic duality and higher derivative $F$ terms. *Nuclear Phys. B* **455** (1995), 109–130.

[5] Bershadsky, M., Cecotti, S., Ooguri, H., Vafa, C., Kodaira-Spencer theory of gravity and exact results for quantum string amplitudes. *Comm. Math. Phys.* **165** (1994), 311–428.

[6] Borcea, C., K3 surfaces with involutions and Mirror Pairs of Calabi-Yau manifolds. In *Mirror Symmetry II*. Ed. by B. Greene and S. T. Yau, AMS/IP Stud. Adv. Math. 1, Amer. Math. Soc./International Press, 1997.

[7] Bryan, J., Pandharipande, R., The local Gromov–Witten theory of curves. Preprint; math.AG/0411037.

[8] Cox, D., Katz, S., *Mirror symmetry and algebraic geometry*. Math. Surveys Monogr. 68, Amer. Math. Soc., Providence, RI, 1999.

[9] Diaconescu, D. E., Florea, B., Grassi, A., Geometric transitions, del Pezzo surfaces and open string instantons. *Adv. Theor. Math. Phys.* **6** (2003), 643–702.

[10] Ferrara, S., Harvey, J. A., Strominger, A., Vafa, C., Second quantized mirror symmetry. *Phys. Lett. B* **361** (1995), 59–65.

[11] Gathmann, A., Topological recursion relations on Gromov-Witten invariants of higher genus. Eprint math.AG/0305361.

[12] Gopakumar, R., Vafa, C., On the gauge theory/geometry correspondence. *Adv. Theor. Math. Phys.* **3** (1999), 1415–1443.

[13] Harvey, J., Moore, G., Algebras, BPS states, and strings. *Nuclear Phys. B* **463** (1996), 315– 368.

[14] Hori, K., Katz, S., Klemm, A., Pandharipande, R., Thomas, R., Vafa, C., Vakil, R., Zaslow, E., *Mirror symmetry*. Clay Math. Monogr. 1, Amer. Math. Soc., Providence, RI, Clay Mathematics Institute, Cambridge, MA, 2003.

[15] Kachru, S., Vafa, C., Exact results for $N = 2$ compactifications of heterotic strings. *Nuclear Phys. B* **450** (1995), 69–89.

[16] Katz, S., Klemm, A., Vafa, C., M-theory, topological strings and spinning black holes. *Adv. Theor. Math. Phys.* **3** (1999), 1445–1537.

[17] Klemm, A., Kreuzer, M., Riegler, E., Scheidegger, E., Topological string amplitudes, complete intersection Calabi-Yau spaces and threshold corrections. *J. High Energy Phys.* **0505** (2005), 023.

[18] Klemm, A., Mariño, M., Counting BPS states on the Enriques Calabi–Yau. Eprint hep-th/0512227.

[19] Li, J., Liu, C.-C. M., Liu, K., Zhou, J., A mathematical theory of the topological vertex. Eprint math.AG/0408426.

[20] Mariño, M., Chern–Simons theory and topological strings. *Rev. Mod. Phys.* **77** (2005), 675–720.

[21] Mariño, M. *Chern–Simons theory, matrix models, and topological strings*. Internat. Ser. Monogr. Phys. 131, Oxford University Press, Oxford 2005.

[22] Mariño, M., Moore, G. W., Counting higher genus curves in a Calabi–Yau manifold. *Nuclear Phys. B* **543** (1999), 592–614.

[23] Maulik, D., Nekrasov, N., Okounkov, A., Pandharipande, R., Gromov–Witten theory and Donaldson–Thomas theory, I and II. Eprints math.AG/0312159 and math.AG/0406092.

[24] Maulik, D., Pandharipande, R., A topological view of Gromov-Witten theory. Eprint math.AG/0412503.

[25] Maulik, D., Pandharipande, R., New calculations in Gromov–Witten theory. Eprint math.AG/0601395.

[26] Ooguri, H., Vafa, C. Knot invariants and topological strings. *Nuclear Phys. B* **577** (2000), 419–438.

[27] Voisin, C., Miroirs et involutions sur les surfaces K3. *Astérisque* **218** (1993) 273–323.

Department of Physics, CERN, Geneva 23, Switzerland and Departamento de Matemática, IST, Lisboa, Portugal

E-mail: marcos@mail.cern.ch

# The Cauchy problem in General Relativity

Igor Rodnianski*

**Abstract.** The paper revisits some of the classical and recent results on the Cauchy problem in General Relativity. Special emphasis is put on the problems concerning existence of a Cauchy development, break-down criteria and stability. The author would like to make a disclaimer that despite its general title the paper is not intended as a comprehensible survey. Due to the space-time constraints many remarkable results and developments are either mentioned briefly or not discussed at all. Most notably this concerns various work on the Einstein equations with matter and symmetry reduced problems.

**Mathematics Subject Classification (2000).** Primary 83C05; Secondary 35L15.

**Keywords.** Einstein equations, existence, break-down, stability.

## 1. Introduction

A mathematical description of General Relativity consists of a $3+1$-dimensional Lorentzian manifold $M$ and a metric $g$ verifying the Einstein equations

$$R_{\alpha\beta}(g) - \frac{1}{2} g_{\alpha\beta} R(g) = 8\pi\, T_{\alpha\beta},$$

where $R_{\alpha\beta}$ and $R$ are respectively the Ricci tensor and scalar curvature of $g$ and $T_{\alpha\beta}$ is the energy-momentum tensor of matter. Among the most popular mathematical matter models are[1]

1. the vacuum equations, where $T \equiv 0$, and the Einstein equations simply require that $(M, g)$ is Ricci flat;
2. the Einstein-scalar field model, where $T_{\alpha\beta} = \partial_\alpha\phi\, \partial_\beta\phi - \frac{1}{2} g_{\alpha\beta}\, \partial^\mu\phi\, \partial_\mu\phi$ and $\phi$ is a real valued scalar field $\phi : M \to R$;
3. the Einstein–Maxwell equations, where $T_{\alpha\beta} = \frac{1}{4\pi}\left(F_\alpha^{\ \mu} F_{\beta\mu} - \frac{1}{4} g_{\alpha\beta} F_{\mu\nu} F^{\mu\nu}\right)$, and $F_{\alpha\beta}$ is the electromagnetic tensor;
4. perfect fluid matter model, where $T_{\alpha\beta} = (\rho + p) u_\alpha\, u_\beta + p\, g_{\alpha\beta}$, and $u^\alpha$ is the four-velocity vector, $p$ is the pressure and $\rho$ is the proper energy density of the fluid.

*The paper was written while the author was visiting the Department of Mathematics at MIT.

[1] In what follows we will use the standard conventions of raising and lowering tensorial indices with the help of the metric $g$ and summing over repeated indices.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The contracted Bianci identity $D^\alpha R_{\alpha\beta} = 2\partial_\beta R$ implies that the gravitational tensor $G_{\alpha\beta} = R_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} R$ is always divergence free, $D^\alpha G_{\alpha\beta} = 0$. As a consequence, evolution equations for the external fields in the models described above follow from the requirement that $D^\alpha T_{\alpha\beta} = 0$. In particular, in the scalar field model $\phi$ must satisfy the wave equation

$$\Box_g \phi = \frac{1}{\sqrt{|g|}} \partial_\alpha \big( g^{\alpha\beta} \sqrt{|g|} \partial_\beta \phi \big) = 0$$

on a curved background $(M, g)$. For the Einstein–Maxwell problem the electromagnetic field obeys the Maxwell equations

$$D^\alpha F_{\alpha\beta} = 0, \quad D_\mu F_{\alpha\beta} + D_\beta F_{\mu\alpha} + D_\alpha F_{\beta\mu} = 0.$$

Finally, for the perfect fluid model,

$$u^\alpha D_\alpha \rho + (\rho + P) D^\alpha u_\alpha = 0, \quad (P + \rho) u^\alpha D_\alpha u_\beta + (g_{\alpha\beta} + u_\alpha u_\beta) D^\alpha P = 0.$$

Mathematical problems in Classical General Relativity can be loosely divided into the following categories:

1. Construction of special solutions (e.g. Minkowski, Schwarzschild, Kerr, Friedman–Robertson–Walker).
2. Mathematics of constraint equations (e.g. construction of solutions, positive mass theorem, Riemannian Penrose inequality).
3. Causality and global properties (e.g. singularity theorems, black hole uniqueness, splitting theorems).
4. Cauchy problem (e.g. existence and uniqueness of solutions, break-down, stability).

The evolution (Cauchy) problem in General Relativity consists of constructing a spacetime $(M, g)$ with the property that for a given data set [2] $(\Sigma_0, g_0, k_0)$, a 3-dimensional Riemannian manifold $\Sigma_0$ with a Riemannian metric $g_0$ and a symmetric 2-tensor $k_0$, there exists an embedding $\Sigma_0 \subset M$ such that $g_0$ coincides with the restriction of $g$ to $\Sigma_0$ and $k_0$ is the second fundamental form of the embedding. Since physically one should not be able to distinguish between different coordinate systems, i.e., the Einstein equations are covariant, a solution of the Cauchy problem can be unique only modulo a diffeomorphism.

The equations are overdetermined and the initial data has to satisfy the constraint equations

$$R_0 - |k_0|^2 + (\operatorname{tr} k_0)^2 = 16\pi T_{00}, \quad \nabla^j k_{0ij} - \nabla_i \operatorname{tr} k_0 = 8\pi T_{0i}, \tag{1}$$

where $R_0$ is the scalar curvature of $g_0$ and $\nabla$ is its Levi-Civita covariant derivative.

[2] For simplicity we describe the Cauchy problem for the vacuum equations. In general, one also needs to add the data for the external fields on $\Sigma_0$.

A particular important class is the asymptotically flat initial data given by $(\Sigma_0, g_0, k_0)$ with the properties[3] that $\Sigma_0$ minus a compact set is diffeomorphic to $R^3$ minus a ball and that there exists a system of coordinates "near infinity" such that

$$g_{0ij} = \left(1 + \frac{M}{r}\right)\delta_{ij} + O(r^{-1-\alpha}), \quad k_{0ij} = O(r^{-2-\alpha})$$

for some $\alpha > 0$. In particular, Minkowski, Schwarzschild and Kerr solutions belong to this class.

## 2. Existence and uniqueness of a maximal Cauchy development

Existence of a maximal Cauchy development for any sufficiently smooth initial data was established by Choquet-Bruhat in [CB1]. The proof exploited the covariance of the Einstein equations and a special choice of gauge (harmonic coordinate system) in which the Einstein equation can be cast as a system of quasilinear wave equations for the metric components.

Harmonic coordinates $x^\alpha$, which appeared already in the work of A.Einstein, are determined by the conditions

$$\Box_g x^\alpha = 0, \quad \alpha = 0, \ldots, 3. \tag{2}$$

This, in turn, is equivalent to requiring that the components of the metric $g_{\alpha\beta}$ relative to this particular system of coordinates satisfy

$$\partial_\alpha\left(g^{\alpha\beta}\sqrt{|g|}\right) = 0. \tag{3}$$

The Einstein equations then reduce to

$$\Box_g g_{\alpha\beta} - N_{\alpha\beta}(g, \partial g) = T_{\alpha\beta}. \tag{4}$$

In harmonic coordinates, $\Box_g = g^{\mu\nu}\partial^2_{\mu\nu}$ and the nonlinear term $N_{\alpha\beta}(u, v)$ depends quadratically on the variable $v$. Equation (4) is obtained by expressing the Ricci tensor $R_{\alpha\beta}$ of $g$ in terms of the components of $g$ and its (first and second) derivatives. To verify that a solution of (4) gives an Einstein metric one also has to satisfy the condition (2) or (3). However, as was observed by Choquet-Bruhat, these conditions are satisfied automatically for solutions of (4) provided that they are satisfied initially on $\Sigma_0$ and $(\Sigma_0, g_0, k_0)$ obey the constraint equations (1). This fact is known as propagation of the harmonic gauge. Thus to complete the initial value problem set up in harmonic gauge we choose a local system of coordinates on $\Sigma_0$ in such a way that the harmonic coordinate condition is verified initially and express the initial values[4]

[3] The asymptotic flatness condition given here is more restrictive than necessary. In particular, it requires that $\Sigma_0$ has only one asymptotic end and that the linear momentum of initial data is equal to zero.

[4] We identify $t = x^0$ and assume that $\Sigma_0$ corresponds to $t = 0$.

for the components of the metric $g_{\mu\nu}|_{t=0}$ and their time derivatives $\partial_t g_{\mu\nu}|_{t=0}$ from the initial data $(g_0, k_0)$[5]. Restricting our analysis for simplicity to the case of the vacuum equations we obtain the following initial value problem

$$\begin{aligned} &\Box_g g_{\alpha\beta} = N_{\alpha\beta}(g, \partial g), \\ &g_{\alpha\beta}|_{t=0} = g^0_{\alpha\beta}, \quad \partial_t g_{\alpha\beta}|_{t=0} = g^1_{\alpha\beta}. \end{aligned} \tag{5}$$

The equations (5) constitute a system of quasilinear wave equations for the components $g_{\alpha\beta}$ on the background determined by the metric $g$. The problem (5) is then solved locally (using finite speed of propagation) on a small interval of time, the resulting metrics patched together to form a Cauchy development from given initial data. The above local solutions are constructed by iteration of the linear equations

$$\begin{aligned} &\Box_{g^n} g^{n+1}_{\alpha\beta} = N_{\alpha\beta}(g^n, \partial g^n), \\ &g^{n+1}_{\alpha\beta}|_{t=0} = g^0_{\alpha\beta}, \quad \partial_t g^{n+1}_{\alpha\beta}|_{t=0} = g^1_{\alpha\beta} \end{aligned} \tag{6}$$

with convergence guaranteed by estimates for (6) or alternatively a priori estimates for (5). The original approach of Choquet-Bruhat to (6) relied on the construction of a Kirchoff–Sobolev parametrix for an inhomogeneous scalar linear problem

$$\Box_g \phi = F, \quad \phi|_{t=0} = \phi^0, \quad \partial_t \phi|_{t=0} = \phi^1$$

but the method imposed high differentiability requirement on the initial data for (5). This was refined in the work of Dionne [Di], Fisher–Marsden [F-M] and Hughes–Kato–Marsden [H-K-M] via the energy method. The energy method, in which the equation (5) is multiplied by $\partial_t g$ and integrated by parts or differentiated required number of times, multiplied by the time derivative of the differentiated solution and then integrated by parts, applied to the problem (5) shows that the Sobolev norm $H^s$ of the solution $g(t)$ at time $t$ is controlled

$$\|g(t)\|_{H^s} + \|\partial_t g(t)\|_{H^{s-1}} \leq C \exp\left(\int_0^t \|\partial g(\tau)\|_{L^\infty} d\tau\right) \left(\|g^0\|_{H^s} + \|g^1\|_{H^{s-1}}\right).$$

The desired a priori estimate for (5) follows from the Sobolev embedding $H^s \subset L^\infty$ provided that $s > 5/2$. This analysis essentially establishes *well-posedness*[6] of the system (5) in the scale of Sobolev spaces $H^s$ for any $s > 5/2$.

An interesting phenomenon however occurs in passage from the system (5) to the original Einstein equations. The above construction leads to solutions $(M, g)$ arising from arbitrary $H^s$ initial data $(g_0, k_0) \in H^s \times H^{s-1}$, as long as $s > \frac{5}{2}$. These solutions remain in the space $H^s$ relative to a system of coordinates $(t, x)$ so that the metric components $g_{\alpha\beta} \in C([0, T]; H^s_x)$ and $\partial_t g_{\alpha\beta} \in C([0, T]; H^{s-1}_x)$ on a time interval $[0, T]$ with $T$ dependent on the $H^s \times H^{s-1}$ norm of the data. However,

[5] Similar procedure is applied to the initial values for the external fields.

[6] Existence, uniqueness and continuous dependence on the initial data.

to show that two solutions $(M, g)$ and $(M', g')$ arising from the same initial data $(\Sigma_0, g_0, k_0)$ are related by a diffeomorphism $\Phi : M \to M'$ so that $\Phi_* g' = g$, and thus geometrically and physically indistinguishable, actually requires one to consider data and thus solutions $(M, g)$ and $(M', g')$ from the Sobolev class $H^\sigma$ with $\sigma > 7/2$. This means that while uniqueness for (5) holds in the same class $H^s$ with $s > 5/2$ as the existence result, there is a potential loss of uniqueness for the Einstein equations unless more regular solutions are considered.

The existence result for the system (5) and consequently the (vacuum or scalar field) Einstein equations can be improved when the energy method is combined with the Strichartz estimates. This was first seen for scalar semilinear wave equations in [P-S],

$$\Box\phi = N(\phi, \partial\phi), \tag{7}$$

where the energy estimate

$$\|\partial\phi(t)\|_{H^s} \leq C \exp\left(\int_0^t \|\phi(\tau)\|_{L^\infty} d\tau\right) \left(\|\phi^0\|_{H^s} + \|\phi^1\|_{H^{s-1}}\right)$$

can be complemented by the Strichartz estimate

$$\|\partial\phi\|_{L^2_{[0,T]}L^\infty} \leq C\left(\|\phi^0\|_{H^s} + \|\phi^1\|_{H^{s-1}} + \|\Box\phi\|_{L^1_{[0,T]}H^{s-1}}\right), \tag{8}$$

which holds for any $s > 2$ and thus allows to establish the existence and uniqueness result for the equation (7) for solutions in the Sobolev space $H^s$ with $s > 2$.

In the case of general quasilinear wave equations of the form (5), however, the situation is far more difficult. One can no longer rely on the Strichartz inequality (8) for the flat D'Alembertian; we need instead its extension to the operator $\Box_g$,

$$\|\partial\phi\|_{L^2_{[0,T]}L^\infty} \leq C\left(\|\phi^0\|_{H^s} + \|\phi^1\|_{H^{s-1}} + \|\Box_g\phi\|_{L^1_{[0,T]}H^{s-1}}\right). \tag{9}$$

To be able to apply such an estimate to the problem (5) and improve upon the energy method one needs to establish (9) for some $s \leq 5/2$ and with a constant $C$ which itself depends on $g$ only through its $\|\partial g\|_{L^\infty_{[0,T]}H^{s-1}}$ and $\|\partial g\|_{L^2_{[0,T]}L^\infty}$ norms. This means that we have to confront the issue of proving Strichartz estimates for wave operators $\Box_g$ on a *rough* background $g$.

This issue was first addressed in the work of Smith [Sm], Bahouri–Chemin [B-C1], [B-C2] and Tataru [Ta1], [Ta2].

In [Sm] a precise analog of (8) was established for the wave operator $\Box_g$ under the assumption that the metric $g$ is at least $C^2$.

The results of Bahouri–Chemin and Tataru are based on establishing a Strichartz type inequality, *with a loss*, i.e. with $s > 2 + \sigma$, and are compatible with applications to the problem (5). The optimal result[7] in this regard, due to Tataru, see [Ta2], requires

[7] Recently Smith-Tataru [S-T1] have shown that the result of Tataru is indeed sharp.

a loss of $\sigma = \frac{1}{6}$. This led to a proof of local well posedness for systems of type (5)[8] with $s > 2 + \frac{1}{6}$.

To do better than that one needs to take into account the nonlinear structure of the equations. Both the classical work [CB1], [Di], [F-M], [H-K-M] and the Strichartz based results [B-C1], [B-C2], [Ta1], [Ta2] only used the fact that the background metric $g$ is Lorentzian and obeys regularity conditions compatible with the final desired result. The additional important piece of information that $g$ itself is a solution of (5) was not exploited.

In [K-R1] we were able to improve the result of Tataru by taking into account not only the expected regularity properties of the metric $g$ but also the fact that they are themselves solutions to a similar system of equations. This allowed us to improve the exponent $s$, needed in the proof of well posedness of equations of type (5) to $s > 2 + \frac{2-\sqrt{3}}{2}$. Our approach was based on a combination of the paradifferential calculus ideas, initiated in [B-C1] and [Ta2], with a geometric treatment of the actual equations introduced in [K4]. The main improvement was due to a gain of conormal differentiability for solutions to the eikonal equations

$$g_{<\lambda}^{\alpha\beta}\partial_\alpha u\,\partial_\beta u = 0 \tag{11}$$

with $g_{<\lambda}$ a smoothed out version of the original metric $g$ with the property that $|\nabla^k g_{<\lambda}| \leq C_k\lambda^k$ for any spatial derivative $\nabla$. Such smoothing can be constructed with the help of the standard Littlewood–Paley projections $P_{<\lambda}$ which smoothly remove Fourier frequencies $\geq \lambda$.

In [K-R2]–[K-R4] we developed the ideas of [K-R1] further in the context of the Einstein vacuum equations, i.e., equations (5) coupled with the condition that $R_{\alpha\beta}(g) = 0$. We make use of both the vanishing of the Ricci curvature of $g$ and the harmonic gauge condition (3). The other important new features are the use of energy estimates along the null hypersurfaces generated by the optical function $u$ and a deeper use of the conormal properties of the null structure equations.

**Theorem 1** ([K-R2]). *Consider the reduced equation* (5) *with initial data* $g^0_{\alpha\beta}, g^1_{\alpha\beta} \in H^s \times H^{s-1}$ *for some* $s > 2$ *satisfying the constraint equations* (1) *with* $T \equiv 0$ *and the harmonic gauge condition* (3). *Then there exists a time interval* $[0, T]$ *and unique* (*Lorentzian metric*) *solution* $g$ *such that* $g_{\alpha\beta} \in C^0([0, T]; H^s)$ *with* $T$ *depending only on the size of* $\|g^0_{\alpha\beta}\|_{H^s} + \|g^1_{\alpha\beta}\|_{H^{s-1}}$.

The results of Theorem 1 require that the initial data can be approximated by a smooth sequence of data satisfying the constraint equations. Using a conformal method a large class of such $(\Sigma_0, g_0, k_0)$ was constructed in [CB3] and [Ma].

[8] The above results actually apply to more general quasilinear equations of the form

$$\Box_{g(\phi)}\phi = N(\phi, \partial\phi), \tag{10}$$

where $g$ is a *given* metric smoothly dependent on a solution $\phi$. However, there is no substantial difference between the equations (5) and (10) unless of course one also uses the fact that a solution of (5) with initial data in harmonic gauge is a solution of the vacuum Einstein equations.

In [S-T2] H. Smith and D. Tataru obtained the parallel $H^s$, $s > 2$ local well posedness result for general quasilinear equations, as well as the new improved results in other dimensions rather than $n = 3$. Their approach is based on the construction of a wave packet approximation of a solution. The geometry of wave packets controls the desired Strichartz estimate. The construction relies on the foliation by the null planes. It uses a gain of differentiablity along each plane, which can be traced to the decomposition of the tangential components of the curvature in the spirit [K-R1], but avoids references to the regularity of the foliation in the direction transversal to the leafs (i.e. torsion of the foliation).

It is very likely that the results of Theorem 1 are not sharp and the Einstein vacuum equations can be solved in even lower degree of regularity. A very satisfactory result both from the analytic and geometric point of view would a resolution of the $L^2$ *curvature conjecture*, see [K3], according to which the time of existence for solutions of the Einstein vacuum equations should depend only on the $L^2$ norms of the Riemann curvature tensor of $g_0$ and the gradient of the second fundamental form $\nabla k_0$ and perhaps some other weaker geometric characteristics of $\Sigma_0$. Some geometric evidence in support of this conjecture is provided in the work [K-R5]–[K-R7] where it was shown that null hypersurfaces, level surfaces of the optical function solving the eikonal equation $g^{\alpha\beta}\partial_\alpha u\,\partial_\beta u = 0$, do not break down locally as long as the $L^2$ curvature flux along them is finite.

**2.1. Existence results in other gauges.** Existence results for the Einstein (vacuum) equations can be also established in other gauges than the harmonic coordinate gauge (2).

In [A-M1] application of the energy method yields a construction of the $H^s$ with $s > 5/2$ vacuum space-times in the constant mean curvature spatially harmonic gauge. To describe the evolution equations in this particular gauge we write the metric $g$ in the form

$$g = -N^2 dt^2 + \gamma_{ij}(dx^i + X^i dt)(dx^j + X^j dt),$$

where $N$ and $X$ are the lapse and shift of the $t$-foliation. The Einstein vacuum equations are written as a system of evolution equations for the metric $\gamma$ and the second fundamental form $k$ of the $t$-foliation coupled to the constraint equations, while the gauge condition generates elliptic equations for $N$ and $X$.

$$\partial_t \gamma_{ij} = -2Nk_{ij} + \mathcal{L}_X \gamma_{ij}, \tag{12}$$

$$\partial_t k_{ij} = -\nabla_i \nabla_j N + N({}^{(3)}R_{ij} + \operatorname{tr} k\, k_{ij} - 2k_{im}k_j^m) + \mathcal{L}_X k_{ij}. \tag{13}$$

Here $\mathcal{L}$ is the Lie derivative and ${}^{(3)}R_{ij}$ is the Ricci curvature of $\gamma$. The constant mean curvature condition is the requirement that on the hypersurface $t = \text{const}$ we have $\operatorname{tr} k = t$. Under this condition taking the trace in (13) and using the constraint equations we obtain an elliptic equation for the lapse $N$:

$$-\Delta_\gamma N + |k|^2 N = 1. \tag{14}$$

The constraint equations in this gauge also become

$$^{(3)}R = |k|^2 - t, \quad \nabla^j k_{ij} = 0. \tag{15}$$

We also fix the spatially harmonic gauge by requiring[9] that a system of coordinates $x^i$, $i = 1, 2, 3$ on each $t =$ constant is harmonic, i.e., satisfies the equation $\Delta_\gamma x^i = 0$. The Ricci curvature $^{(3)}R_{ij}$ can then be written on the form

$$^{(3)}R_{ij} = -\frac{1}{2}\Delta_\gamma \gamma_{ij} + N_{ij}(\gamma, \nabla\gamma),$$

where as before $N_{ij}(u, v)$ depends quadratically on $v$. Propagation of this gauge results in an elliptic equation for the shift so that (12), (13), (14), (15) form an elliptic-hyperbolic system.

A local existence result in the maximal gauge was also proved in [C-K]. This particular gauge corresponds to the choice $\operatorname{tr} k = X = 0$. The lapse and constraint equations take the form

$$-\Delta_\gamma N + |k|^2 N = 0,$$
$$^{(3)}R = |k|^2, \quad \nabla^j k_{ij} = 0$$

while the system (12), (13) describes the evolution of $\gamma$ and $k$. To see the hyperbolic character of (12), (13) without imposing a spatially harmonic gauge one has to take an additional time derivative of (13) and express $\partial_t^{(3)}R$ in terms of $\gamma$ and $k$.

Another interesting formulation of the Einstein (vacuum) equations arises by drawing an analogy between the Einstein equations and the Yang–Mills theory. In the Yang–Mills theory an electromagnetic field is represented by a Lie algebra valued 2-form $F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha + [A_\alpha, A_\beta]$ constructed from an electromagnetic potential $A_\alpha$. The Yang–Mills equations on Minkowski space $R^{3+1}$ for $F$ take the form

$$D^\beta F_{\alpha\beta} = 0, \quad D_\mu F_{\alpha\beta} + D_\beta F_{\mu\alpha} + D_\alpha F_{\beta\mu} = 0,$$

where the covariant derivative $D_\alpha = \partial_\alpha + [A_\alpha, \cdot]$ and the second equation is the Bianci identity for the curvature form $F$. Differentiating the second equation we arrive at the second order hyperbolic problem for $F$

$$\Box_A F_{\alpha\beta} = 2F_\beta^{\ \mu} F_{\mu\alpha},$$

where $\Box_A = m^{\alpha\beta} D_\alpha D_\beta$.

In General Relativity the Riemann curvature tensor $R_{\alpha\beta\mu\nu}$ satisfies the Bianci identities

$$D_\sigma R_{\alpha\beta\mu\nu} + D_\beta R_{\sigma\alpha\mu\nu} + D_\alpha R_{\beta\sigma\mu\nu} = 0 \tag{16}$$

[9]The actual general harmonic gauge is only slightly more complicated.

and, if the Ricci curvature of $g$ vanishes, i.e., $(M, g)$ is a vacuum space-time, also a version of the contracted Bianci identities

$$D^{\alpha} R_{\alpha\beta\mu\nu} = 0. \tag{17}$$

Differentiating the Bianci identity (16) and also using (17) we easily obtain the wave equation for the Riemann curvature tensor

$$\Box_g R_{\alpha\beta\mu\nu} = (R \star R)_{\alpha\beta\mu\nu}\,, \tag{18}$$

where $\star$ denotes a combination of various contractions.

**2.2. Large data problem in General Relativity.** While the result of Choquet-Bruhat and its subsequent refinements guarantee the existence and uniqueness of a (maximal) Cauchy development, they provide no information about its geodesic completeness and thus, in the language of partial differential equations, constitutes a local existence result. Singularities could develop to the future (past) of the Cauchy hypersurface $\Sigma_0$ or the maximal Cauchy development could have a regular boundary, Cauchy horizon, beyond which the space-time could be continued thus losing its predictability from the initial data. Schwarzschild space-time is an example of a geodesically incomplete asymptotically flat space-time while the Reissner–Nordström solution of the Einstein–Maxwell equations possesses a Cauchy horizon.

More generally, there are a number of conditions that will guarantee that the space-time will be geodesically incomplete. The first such result was the Penrose singularity theorem:

**Theorem 2** (Penrose). *A space-time*[10] $(M, g)$ *is necessarily incomplete if it admits a non-compact Cauchy surface and a trapped*[11] 2*-dimensional compact surface.*

According to the result of [S-Y2] sufficient amount of matter placed in a region will create a trapped surface.

In the language of partial differential equations this means an impossibility of a large data global existence result for all initial data in General Relativity. In the absence of such a result a number of conjectures about the structure of space-times arising from generic data had been put forward in the '60s by Penrose. Among them is the *Weak Cosmic Censorship* which predicts that for generic asymptotically flat data null infinity will be affine complete or alternatively that singularities have to be hidden inside black holes. On the other hand the *Strong Cosmic Censorship* predicts a generic absence of the Cauchy horizons. At the moment not much progress has been made on either of these problems in the general case with remarkable exceptions in some cases of symmetry reduced Einstein equations: the proof of weak

[10] The energy-momentum tensor $T_{\alpha\beta}$ of matter is only required to satisfy what is called a null convergence energy condition.

[11] Infinitesimally deformations of such surface along both null outgoing and null incoming directions decrease its area.

cosmic censorship for the Einstein-scalar field equations in spherical symmetry by Christodoulou [C2],[C3], the work of Dafermos [D] on strong cosmic censorship and stability of the Cauchy horizons for the Einstein–Maxwell-scalar field equations in spherical symmetry, the proof of strong cosmic censorship in polarized Gowdy by Chruściel–Isenberg–Moncrief [C-I-M] and $T^3$-Gowdy by Ringström [Ri].

**2.3. Break-down criteria in General Relativity.** In the absence of a completeness result for general large data Cauchy problem in General Relativity and a very non-quantitative nature of the singularity theorems it is desirable to develop a better understanding of local or semi-local analytic mechanisms for break-down of solutions. Already the local existence results mentioned above provide such criteria. In particular an $H^s$ local existence result in harmonic gauge guarantees that a solution can be extended as long as the $H^s$ norms of the metric components in harmonic gauge remain finite. Such results however are not ideal as break-down criteria for the reasons that they are not geometric and strongly tied to a particular coordinate gauge and that they arise as a consequence of stronger local well-posedness statements.

The first *geometric* criterion for breakdown of solutions $(M, g)$ of the vacuum Einstein equations

$$R_{\alpha\beta}(g) = 0 \tag{19}$$

appeared in the work of M. Anderson [A1]. To describe the problem we assume that a part of space-time $M_I \subset M$ is foliated by the level hypersurface $\Sigma_t$ of a time function $t$, monotonically increasing towards future in the interval $I \subset R$, with lapse $N$ and second fundamental form $k$ so that

$$g = -N^2 dt^2 + \gamma_{ij} dx^i dx^j, \quad \partial_t \gamma_{ij} = -2N k_{ij}.$$

The surfaces $\Sigma_t$ are compact, of Yamabe type $-1$, and of constant negative mean curvature, $\operatorname{tr} k = t$ with $t < 0$. Relative to a time foliation we can naturally associate a non-degenerate notion of a pointwise absolute value of a space-time tensor.

In [A1] it was shown that a break-down can be tied to the condition that

$$\limsup_{t \to t_*^-} \|R(t)\|_{L^\infty} = \infty,$$

where $R(t)$ denotes the Riemann curvature tensor of $g$ and the norm is measured relative to the above described $t$-foliation.

A work in progress [K-R7] addresses the problem of break-down of solutions to the Einstein vacuum equations under the assumption that $T = N^{-1}\partial_t$ is an approximate Killing field. More precisely, the desirable break-down condition is

$$\limsup_{t \to t_*^-} \|\mathcal{L}_T g(t)\|_{L^\infty} = \infty, \tag{20}$$

where $\mathcal{L}_T g$ is the deformation tensor of $T$, equal to zero if $T$ is Killing, and it can be expressed as

$$|\mathcal{L}_T g| = |k| + |\nabla \log N|.$$

This result would complement Anderson's criterion. It is clear however that the condition (20) is formally weaker as it refers only to the second fundamental form $k$ and the lapse $n$ and thus requires one degree less of differentiability than a condition on the Riemann curvature tensor. Moreover a condition on the boundedness of the $L^\infty$ norm of $R(t)$ covers all the dynamical degrees of freedom of the equations. Indeed, once we know that $\|R(t)\|_{L^\infty}$ is finite, one can find bounds for $n$, $\nabla n$ and $k$ on $\Sigma_t$ purely by elliptic estimates. This is not true in our case.

A geometric criterion of the type (20) for the Einstein equations could be compared to the well known Beale–Kato–Majda [B-K-M] criterion for breakdown of solutions of the incompressible Euler equation

$$\partial_t v + (v \cdot \nabla)v = -\nabla p, \quad \operatorname{div} v = 0,$$

with smooth initial data at $t = t_0$. A routine application of the energy estimates shows that solution $v$ blows up if and only if

$$\int_{t_0}^{t_*} \|\nabla v(t)\|_{L^\infty}\, dt = \infty. \tag{21}$$

The Beale–Kato–Majda work improves the blow up criterion by replacing it with the following condition on the vorticity $\omega = \operatorname{curl} v$:

$$\int_{t_0}^{t_*} \|\omega(t)\|_{L^\infty}\, dt = \infty. \tag{22}$$

The quantities $\nabla v$ and $\omega$ are related to each other via a singular integral operator, i.e., $\nabla v = P^0(\omega)$.

Although $P^0$ does not define a bounded map $L^\infty \to L^\infty$ it is sufficient to reduce the breakdown condition (21) to the more satisfying one (22), in terms of the vorticity alone.

Similarly, in the case of the Einstein equations energy estimates, expressed relative to a special system of coordinates (e.g. in harmonic gauge), show that break-down does not occur unless

$$\int_{t_0}^{t_*} \|\partial g(t)\|_{L^\infty}\, dt = \infty.$$

This condition however is not geometric as it depends on the choice of a full coordinate system. Observe that both the spatial derivatives of the lapse $\nabla n$ and the components of the second fundamental form, $k_{ij} = -\frac{1}{2}N^{-1}\,\partial_t g_{ij}$, can be interpreted as components of $\partial g$.

Note however that after prescribing $k$ and $\nabla n$ we are still left with many more degrees of freedom in determining $\partial g$. The fundamental difficulty that one needs to overcome is that of deriving bounds for $R$ using only bounds for $\|\nabla N(t)\|_{L^\infty} + \|k(t)\|_{L^\infty}$ and geometric informations on the initial hypersurface $\Sigma_0$. Clearly this cannot be done by elliptic estimates alone. Thus, as opposed to both the results of M.

Anderson and Beale–Kato–Majda, it is far less obvious that a condition such as (20) can cover all *dynamic* degrees of freedom of the Einstein equations.

The criterion (20) is motivated in part by the desire to adapt the Eardley–Moncrief argument [E-M1], [E-M2] for the large data global existence for the $3+1$ Yang–Mills equations to General Relativity, exploiting the analogy between the Einstein vacuum and the Yang–Mills equations.

The Eardley–Moncrief proof relies on two independent ingredients: conservation of energy and pointwise bounds on curvature, which are derived using the fundamental solution for $\Box$ in Minkowski space, and shown to depend only on the flux of curvature and initial data. Since the analog of the Yang–Mills energy in General Relativity (the Bel–Robinson energy) is not conserved one can only hope to reproduce the second part of the Eardley–Moncrief argument and prove a conditional regularity result which states, roughly, that smooth solutions of the Einstein equations, in vacuum, remain smooth, and can therefore be continued, as long as an integral quantity, we call the flux of curvature, remains bounded. The possibility of such a result was first pointed out by V. Moncrief.

However it is the fact that $T = \partial_t$ is a Killing field that is ultimately responsible for the conservation of energy in the Yang–Mills theory on Minkowski space. Similarly, in the extension [C-S] of the Eardley–Moncrief result to the Yang–Mills equations on a smooth globally hyperbolic background it is the fact that $T = \partial_t$ is an approximate Killing field that allows one to control the energy and the flux of curvature.

Thus in the context of General Relativity rather then imposing a direct condition on the finiteness of the Bel–Robinson energy and curvature flux we formulate conditions (perhaps more natural albeit more restrictive) which control the extent to which the energy is not conserved. These conditions, which form our breakdown criterion, involve uniform bounds on the second fundamental form $k$ and the lapse $N$.

In what follows we describe how the main ideas of the proof of the Eardley–Moncrief result for Yang–Mills could be adapted to General Relativity.

The curvature tensor $R$ of a $3+1$ dimensional vacuum spacetime $(M, g)$, see (19), verifies a wave equation of the form,

$$\Box_g R = R \star R. \tag{23}$$

The Bel–Robinson energy-momentum tensor

$$\mathcal{Q}[R]_{\alpha\beta\gamma\delta} = R_{\alpha\lambda\gamma\mu} R_{\beta\ \delta}^{\ \lambda\ \mu} + \star R_{\alpha\lambda\gamma\mu} \star R_{\beta\ \delta}^{\ \lambda\ \mu}$$

verifies $D^\delta \mathcal{Q}[R]_{\alpha\beta\gamma\delta} = 0$ and can thus be used to derive energy and flux estimates for the curvature tensor $R$. The approximate Killing condition is sufficient to derive bounds for both energy and flux associated to the curvature tensor $R$. The flux is an integral of a square of the components of the Riemann curvature tensor tangent to a null hypersurface $N^-(p)$, boundary of the causal point of point $p$, generated, at least locally, as a level hypersurface $u = 0$ of an optical function $u$, solution of the eikonal equation $g^{\alpha\beta}\partial_\alpha u\, \partial_\beta u = 0$.

As in the case of the Yang–Mills equations it is precisely the boundedness of the flux of curvature that plays a crucial role in our analysis. In General Relativity the flux takes on even more fundamental role as it is also needed to control the geometry of the very object it is defined on, i.e. the boundary $N^-(p)$ of the causal past of $p$. This boundary, unlike in the case of Minkowski space, are not determined a-priori but depend in fact on the space-time we are trying to control.

The main idea is to show that if the condition (20) does not hold, i.e.,

$$\limsup_{t \to t_*^-} \|\mathcal{L}_T g(t)\|_{L^\infty} < \infty, \tag{24}$$

it implies a uniform curvature bound

$$\limsup_{t \to t_*^-} \|R(t)\|_{L^\infty} < \infty \tag{25}$$

and the solution can be continued beyond $t_*$.

The curvature bound (25) relies on the parametrix construction for the equation (23). In the construction of a parametrix for (23) we cannot, in any meaningful way, approximate $\Box_g$ by the flat D'Alembertian $\Box$. One could instead proceed via a geometric optics construction of parametrices for $\Box_g$, as developed in [F]. Such an approach would require additional bounds on the background geometry, determined by the metric $g$, incompatible with the assumption (24) and the implied finiteness of the curvature flux.

We rely instead on a geometric version, which we develop in [K-R6], of the Kirchoff–Sobolev formula, similar to that used by Sobolev in [Sob] and Choquet-Bruhat in [CB1], see also [Mo]. Roughly, this can be obtained by applying to (23) the measure $A\delta(u)$, where $u$ is an optical function whose level set $u = 0$ coincides with $N^-(p)$ and $A$ is a 4-covariant 4-contravariant tensor defined as a solution of a transport equation along $N^-(p)$ with appropriate (blowing-up) initial data at the vertex $p$. After a careful integration by parts we arrive at the following analogue of the Kirchoff formula:

$$R(p) = -\int_{N^-(p;\delta_0)} A \cdot (R \star R) + R^0(p;\delta_0) + \int_{N^-(p;\delta_0)} \mathcal{E} \cdot R, \tag{26}$$

where $N^-(p;\delta_0)$ denotes the portion of the null boundary $N^-(p)$ in the time interval $[t(p) - \delta_0, t(p)]$ and the error term $\mathcal{E}$ depends only on the intrinsic geometry of $N^-(p;\delta_0)$. The term $R^0(p;\delta_0)$ is completely determined by the initial data on the hypersurface $\Sigma_{t(p)-\delta_0}$. As in the flat case[12], one can prove bounds for the sup-norm of $R^0(p;\delta_0)$ which depend only on uniform bounds for $R$ and its first covariant derivatives at values of $t' \le t(p) \le t - \delta$.

As in the Yang–Mills setting the structure of the term $R \star R$ allows us to estimate one of the curvature terms by the flux of curvature.

[12]This is by no means obvious as we need to rely once more on the Kirchoff–Sobolev formula.

To control the error term in (26) one needs estimates for tangential derivatives of $A$ and other geometric quantities associated to the null hypersurfaces $N^-(p)$. In particular, it requires showing that $N^-(p)$ remains a *smooth* (not merely Lipschitz) hypersurface in the time slab $(t(p) - \delta_0, t(p)]$ for some $\delta_0$ dependent only on the initial data and (24). Thus to prove the desired theorem one would have to show that all geometric quantities, arising in the parametrix construction, can be estimated only in terms of the flux of the curvature along $N^-(p)$ and the bound in (24). Yet, to start with, it is not even clear that we can provide a lower bound for the radius of injectivity of $N^-(p)$. In other words the congruence of null geodesics, initiating at $p$, may not be controllable[13] only in terms of the curvature flux. Typically, in fact, lower bounds for the radius of conjugacy of a null hypersurface in a Lorentzian manifold are only available in terms of the sup-norm of the curvature tensor $R$ along the hypersurface, while the problem of short, intersecting, null geodesics appears not to be fully understood even in that context. The situation is similar to that in Riemannian geometry, exemplified by the Cheeger's theorem, where pointwise bounds on sectional curvature are sufficient to control the radius of conjugacy but to prevent the occurrence of short geodesic loops one needs to assume in addition an upper bound on the diameter and a lower bound on the volume of the manifold.

In a sequence of papers, [K-R2]–[K-R4], we have been able to prove a lower bound, depending essentially only on the curvature flux, for the radius of conjugacy of null hypersurfaces[14] in a Lorentzian spacetime which verifies the Einstein vacuum equations. The methods used in these articles can be adapted to provide all the desired estimates, except a lower bound on the "size" of intersecting null geodesics which needs a separate argument. The lower bound on the radius of injectivity of the null hypersurfaces $N^-(p)$ has been established in [K-R5]

## 3. Stability problems in General Relativity

In the absence of a general "large data" result in General Relativity the problem of stability of special solutions becomes simultaneously more important and more tractable. Despite our best efforts however these stability questions appear to be quite difficult and still poorly understood. A singular achievement in this regard has been the proof of stability of Minkowski space-time, [C-K], [L-R2] and a semi-global version in [Fr1]. To this date this is the only global result in the category of the asymptotically flat space-times. In the realm of cosmological models stability of the de Sitter space has been shown in [Fr2], [A2]. Finally one should also mention the proof of stability in the expanding direction of a flat cone solution for spatially compact space-times [A-M2], [Re].

[13] Different null geodesics of the congruence may intersect, or the congruence itself may have conjugate points, arbitrarily close to $p$.

[14] together with many other estimates of various geometric quantities associated to $N^-(p)$.

**3.1. Stability of the Minkowski space-time.** The problem of stability of Minkowski space-time for the Einstein-vacuum equations can be described as follows:

*Show existence of a causally geodesically complete vacuum space-time asymptotically "converging" to the Minkowski space-time for an arbitrary set of smooth asymptotically flat initial data* $(\Sigma_0, g_{0ij}, k_{0ij})$ *with* $\Sigma_0 \approx R^3$,

$$g_{0ij} = \left(1 + \frac{M}{r}\right)\delta_{ij} + o(r^{-1-\alpha}), \quad k_{0ij} = o(r^{-2-\alpha}), \quad r = |x| \to \infty,\ \alpha > 0 \tag{27}$$

*where* $(g_0 - \delta)$ *and* $k_0$ *satisfy global smallness assumptions .*

A positive parameter $M$ in the asymptotic expansion for the metric $g_0$ is the ADM mass, positive according to [S-Y1], [W].

The stability of Minkowski space for the Einstein-vacuum equations was shown in a remarkable work of Christodouolou–Klainerman for strongly asymptotic initial data (the parameter $\alpha \geq 1/2$ in the asymptotic expansion (27)), [C-K] The approach taken in that work viewed the Einstein-vacuum equations as a system of equations

$$D^\alpha W_{\alpha\beta\gamma\delta} = 0, \quad D^\alpha * W_{\alpha\beta\gamma\delta} = 0$$

for the Weyl tensor $W_{\alpha\beta\gamma\delta}$ of the metric $g_{\alpha\beta}$ and used generalized energy inequalities associated with the Bel–Robinson energy-momentum tensor, constructed from components of $W$, and special geometrically constructed vector fields, designed to mimic the rotation and the conformal Morawetz vector fields of the Minkowski space-time, i.e., "almost conformally Killing" vector fields of the unknown metric $g$. The proof was manifestly invariant, in particular it did not use the harmonic coordinate gauge. This approach was later extended to the Einstein–Maxwell equations by N. Zipser [Z].

In [L-R2] we succeeded in developing a new relatively technically simple approach which allowed allowing us to prove stability of Minkowski space in harmonic coordinate gauge, for general asymptotically flat data, $\alpha > 0$, and simultaneously treat the case of the Einstein equations coupled to a scalar field,

$$R_{\alpha\beta}(g) = \partial_\alpha\phi\, \partial_\beta\phi, \quad \Box_g\phi = 0$$

where the scalar field requires a global smallness assumption on its initial data $(\phi_0, \phi_1)$, which obey the asymptotic expansion

$$\phi_0 = o(r^{-1-\alpha}), \quad \phi_1 = o(r^{-2-\alpha}). \tag{28}$$

**Theorem 3** ([L-R2]). *Let* $(\Sigma, g_0, k_0, \phi_0, \phi_1)$ *be initial data for the Einstein-scalar field equations. Assume that the initial time slice* $\Sigma$ *is diffeomorphic to* $R^3$ *and admits a global coordinate chart relative to which the data is close to the initial data for the Minkowski space-time. More precisely, we assume that the data* $(g_0, k_0, \phi_0, \phi_1)$ *is smooth asymptotically flat in the sense of* (27)–(28) *with mass* $M$ *and* $\alpha > 0$ *and satisfy a global smallness assumption as measured in the scale of weighted Sobolev spaces. Then the Einstein-scalar field equations possess a future causally geodesically*

*complete solution* $(g, \psi)$ *asymptotically converging to Minkowski space-time. In fact, there exists a global harmonic system coordinates relative to which the metric g remains close* (*and "converges"*) *to the Minkowski metric.*

The appeal of the harmonic gauge for the proof of stability of Minkowski space-time lies in the fact that the latter can be simply viewed[15] as a *small data global existence result* for the quasilinear system (5) (for vacuum equations),

$$\Box_g g_{\alpha\beta} = N_{\alpha\beta}(g, \partial g), \quad g_{\alpha\beta}|_{t=0} = g^0_{\alpha\beta}, \ \partial_t g_{\alpha\beta}|_{t=0} = g^1_{\alpha\beta}. \tag{29}$$

However, usefulness of the harmonic gauge in this context was questioned earlier and it was suspected that harmonic coordinates are "unstable in the large", [CB1]. The conclusion is suggested from the analysis of the iteration scheme for the system (29), which resembles an iteration scheme for the semilinear equation $\Box\phi = (\partial_t\phi)^2$ shown to blow up in finite time for arbitrarily small initial data by F. John, [J].

To describe some of the difficulties in establishing a small data global existence result for the system (29) consider a generic quasilinear system of the form

$$\Box\,\phi_i = \sum b_i^{jk\alpha\beta}\,\partial_\alpha\phi_j\,\partial_\beta\phi_k + \sum c_i^{jk\alpha\beta}\,\phi_j\,\partial_\alpha\partial_\beta\phi_k + \text{cubic terms}. \tag{30}$$

The influence of cubic terms is negligible while the quadratic terms are of two types, the *semilinear terms* and the *quasilinear terms*, each of which present their own problems. D. Christodoulou [C1] and S. Klainerman [K2] showed global existence for systems of the form (30) if the semilinear terms satisfy the *null condition* and the quasilinear terms are absent. The null condition, first introduced by S. Klainerman in [K1], was designed to detect systems for which solutions are asymptotically free and decay like solutions of a linear equation. It requires special algebraic cancellations in the coefficients $b_i^{jk\alpha\beta}$, e.g. $\Box\,\phi = (\partial_t\phi)^2 - |\nabla_x\phi|^2$. However, the semilinear terms for the Einstein equations do not satisfy the null condition, see [CB2]. The quasilinear terms is another source of trouble. The only non-trivial example of a quasilinear equation of the type (30), for which the small data global existence result holds, is the model equation $\Box\,\phi = \phi\,\Delta\phi$, as shown in [L1] (radial case) and [Al] (general case), see also [L2].

In [L-R1] we identified a criteria under which it is more likely that a quasilinear system of the form (30) has global solutions[16]. We said that a system of the form (30) satisfy the *weak null condition* if the corresponding *asymptotic system* (c.f. [H]) has global solutions. We showed that the Einstein equations in harmonic coordinates satisfy the weak null condition. In addition an additional cancellation mechanism was found for the Einstein equations in harmonic coordinates that makes it better than a

[15]This statement requires additional care since a priori there is no guarantee that obtained "global in time" solution $g_{\mu\nu}$ defines a causally geodesically complete metric. However, the latter can be established provided one has good control on the difference between $g_{\mu\nu}$ and the Minkowski metric $m_{\mu\nu}$.

[16]At this point, it is unclear whether this criteria is sufficient for establishing a "small data global existence" result for a *general* system of quasilinear hyperbolic equations.

general system satisfying the weak null condition. The system decouples to leading order, when decomposed relative to the Minkowski *null frame*. An approximate model that describes the semilinear terms has the form

$$\Box\,\phi_2 = (\partial_t \phi_1)^2, \quad \Box\,\phi_1 = 0.$$

While every solution of this system is global in time, the system fails to satisfy the classical null condition and solutions are not asymptotically free: $\phi_2 \sim \varepsilon t^{-1} \ln |t|$. The semilinear terms in Einstein's equations can be shown to either satisfy the classical null condition or decouple in the above fashion when expressed in a null frame. The quasilinear terms also decouple but in a more subtle way. The influence of quasilinear terms can be detected via asymptotic behavior of the characteristic surfaces of metric $g$. It turns out that the main features of the characteristic surfaces at infinity are determined by a particular *null* component of the metric. The asymptotic flatness of the initial data and the harmonic coordinate condition (3)

$$\partial_\beta\big(g^{\alpha\beta}\sqrt{|g|}\big) = 0 \tag{31}$$

give good control of this particular component, i.e., $\sim M/r$, which in turn implies that the light cones associated with the metric $g$ diverge only logarithmically $\sim M \ln t$ from the Minkowski cones. The main simplification in our approach comes from the fact the behavior of the system (29) coupled to the harmonic gauge (31) can be completely controlled by means of the generalized energy estimates exploiting only the *exact* symmetries of Minkowski space thus avoiding having to construct dynamically generators of the approximate symmetries of the space-time $(M, g)$.

The asymptotic behavior of null components of the Riemann curvature tensor $R_{\alpha\beta\gamma\delta}$ of metric $g$ – the so called "peeling estimates" – was discussed in the works of Bondi, Sachs and Penrose and becomes important in the framework of asymptotically simple space-times (roughly speaking, space-times which can be conformally compactified), see also the paper of Christodoulou [C4] for further discussion of such space-times. The work of [C-K] provided very precise, although not entirely consistent with peeling estimates, analysis of the asymptotic behavior of constructed global solutions. However, global solutions obtained by Klainerman–Nicolo [K-N1] in the problem of exterior[17] stability of Minkowski space were shown to possess peeling estimates for special initial data, [K-N2]. The work [L-R2] is less precise about the asymptotic behavior of the curvature components.

**3.2. Beyond stability of Minkowski space-time.** The simplest solutions of the Einstein vacuum equations of general relativity,

$$R_{\mu\nu} = 0, \tag{32}$$

[17] Outside of the domain of dependence of a compact set.

containing black holes are the one-parameter Schwarzschild family of solutions. In the exterior region ($r > 2M$) the Schwarzschild metric can be written in the form

$$g_s = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\sigma_{S^2}.$$

The Schwarzschild family is a sub-family of the two-parameter Kerr family which describe stationary rotating black holes. In its proper rigorous formulation, the problem of nonlinear stability of the Kerr family is one of the major open problems in general relativity. In particular, it is conjectured that perturbations of Schwarzschild initial data should evolve into a spacetime with complete null infinity whose past "suitably" approaches a nearby Kerr exterior. At the heuristic level, however, considerable progress has been made in the last 40 years towards an understanding of the issues involved. In particular, a very influential role was played by the work of R. Price [Pr] in 1972, who discovered a heuristic mechanism, known in the physics literature as the *red-shift effect*, allowing for the decay of scalar field linear perturbations on the Schwarzschild exterior, i.e., solutions of the linear wave equation

$$\Box_{g_s} \phi = 0.$$

Despite the abundance of heuristic and numerical arguments the nonlinear problem is still lacking proper mathematical understanding while some progress has been made recently on a problem of asymptotic behavior of the linear problem. The causal picture of the Schwarzschild space-time is very different from the one of the Minkowski space. In (a right quadrant of) Schwarzschild space-time in addition to the null infinity, parametrized by $(u, \omega) \in R \times S^2$ there is a special null hypersurface, the *event horizon*, parametrized by $(v, \omega) \in R \times S^2$, separating the exterior region from the black hole. It is the presence of the event horizon that is responsible for the red-shift effect in which the frequency of an observer leaving the exterior region gets shifted to the red as viewed by the second observer positioned to the future of the first one. The geometry of the Schwarzschild space-time also ultimately determines the behavior of linear waves. However, even uniform boundedness of solutions of linear scalar wave equations, almost trivial in Minkowski context, is by no means obvious and is termed as linear stability of Schwarzschild in the physics literature. This result was rigorously established in the work of Kay and Wald [K-W]. Decay for $\phi$, without a rate, was first proven in [Tw].

**Theorem 4** ([D-R2]). *Let $\phi$ be a sufficiently regular solution of the wave equation*

$$\Box_{g_s} \phi = 0 \tag{33}$$

*on the (maximally extended) Schwarzschild spacetime $(\mathcal{M}, g)$, decaying suitably at spatial infinity on an arbitrary complete asymptotically flat Cauchy surface $\Sigma$. Fix retarded and advanced Eddington–Finkelstein coordinates $u$ and $v$. We have the*

*following pointwise decay rates*

$$\begin{aligned} |\phi| &\leq C \max(1, v)^{-1} \quad \text{in } r \geq 2M \\ |r\phi| &\leq C_{\hat{R}}(1 + |u|)^{-\frac{1}{2}} \quad \text{in } \{r \geq \hat{R} > 2M\} \cap J^+(\Sigma). \end{aligned} \tag{34}$$

A variant of the problem considered here is also studied in [B-S].

In the spherically symmetric case, the above result follows from a very special case of [D-R1], where the so called Price law has been established. (See also [M-S].)

For the more general Kerr family, even uniform boundedness remains an open problem (see however [FKSY]).

The proof of Theorem (4) is based on the energy type estimates for (33) with vector fields adapted to different regions of space-times. An important role in this analysis is played by the "red-shift vector field", which has no equivalent in Minkowski space, constructed near the event horizon.

## References


[Al] Alinhac, S., An example of blowup at infinity for a quasilinear wave equation. *Astérisque* **284** (2003), 1–91.

[A1] Anderson, M., On long-time evolution in general relativity and geometrization of 3-manifolds. *Comm. Math. Phys.* **222** (2001), 533–567.

[A2] Anderson, M., Existence and stability of even dimensional asymptotically de Sitter spaces. *Ann. Henri Poincaré* **6** (2005), 801–820.

[A-M1] Andersson, L., and Moncrief, V., Elliptic-hyperbolic systems and the Einstein equations. *Ann. Henri Poincaré* **4** (2003), 1–34.

[A-M2] Andersson, L., and Moncrief, V., Future complete vacuum spacetimes. In *The Einstein Equations and the Large Scale Behavior of Gravitational Fields: 50 Years of the Cauchy Problem in General Relativity* (ed. by P. T. Chruściel and H. Friedrich), Birkhäuser, Basel 2004, 299–330.

[B-C1] Bahouri, H., and Chemin, J. Y., Équations d'ondes quasilinéaires et estimation de Strichartz. *Amer. J. Math.* **121** (1999), 1337–1777.

[B-C2] Bahouri, H., and Chemin, J. Y., Équations d'ondes quasilinéaires et effet dispersif. *Internat. Math. Res. Notices* **1999** (21) (1999), 1141–1178.

[B-K-M] Beale, T., Kato, T., Majda, A., Remarks on the breakdown of smooth solutions for the 3-D Euler equations. *Comm. Math. Phys.* **94** (1984), 61–66.

[B-S] Blue, P., and Sterbenz, J., Uniform decay of local energy and the semi-linear wave equation on Schwarzchild space. Preprint.

[CB1] Choquét-Bruhat, Y., Théorème d'existence pour certains systèmes d'equations aux dérivées partielles nonlinéaires. *Acta Math.* **88** (1952), 141–225.

[CB2] Choquét-Bruhat, Y., Un théorème d'instabilité pour certaines équations hyperboliques non linéaires. *C. R. Acad. Sci. Paris Ser. A* **276** (1973), 281–284.

[CB3] Choquét-Bruhat, Y., Einstein constraints on $n$ dimensional compact manifolds. *Classical Quantum Gravity* **21** (2004), S127–S152.

[C1] Christodoulou, D., Global solutions of nonlinear hyperbolic equations for small initial data. *Comm. Pure Appl. Math.* **39** (1986), 267–282.

[C2] Christodoulou, D., Bounded variation solutions of the spherically symmetric Einstein-scalar field equations. *Comm. Pure Appl. Math.* **46** (1993), 1131–1220.

[C3] Christodoulou, D., Instability of naked singularities in the gravitational collapse of a scalar field. *Ann. of Math.* **149** (1999), 183–217.

[C4] Christodoulou, D., The Global Initial Value Problem in General Relativity. In *The Ninth Marcel Grossmann Meeting* (Rome, 2000), ed. by V. G. Gurzadyan et al., World Scientific, Singapore 2002, 44–54.

[C-K] Christodoulou, D., Klainerman, S., *The global nonlinear stability of the Minkowski space*, Princeton Math. Ser. 41, Princeton University Press, Princeton, NJ, 1993.

[C-I-M] Chruściel, P. T., Isenberg, J., Moncrief, V., Strong cosmic censorship in polarized Gowdy spacetimes. *Classical Quantum Gravity* **7** (1990), 1671–1680.

[C-S] Chrúsciel, P., Shatah, J., Global existence of solutions of the Yang-Mills equations on globally hyperbolic four-dimensional Lorentzian manifolds. *Asian J. Math.* **1** (3) (1997), 530–548.

[D] Dafermos, M., Stability and instability of the Cauchy horizon for the spherically symmetric Einstein-Maxwell-scalar field equations. *Ann. of Math.* **158** (3) (2003), 875–928.

[D-R1] Dafermos, M., and Rodnianski, I., A proof of Price's law for the collapse of a self-gravitating scalar field. *Invent. Math.* **162** (2005), 381–457.

[D-R2] Dafermos, M., and Rodnianski, I., The red-shift effect and radiation decay on black hole spacetimes. arXiv gr-qc/0512119.

[Di] Dionne, P. A., Sur les problèmes de Cauchy hyperbolique bien posés. *J. Analyse Math.* **10** (1962), 1–90.

[HE] Hawking, S. W., and Ellis, G. F. R., *The Large Scale Structure of Space-time*. Cambridge Monogr. Math. Phys. 1, Cambridge University Press, London, New York 1973.

[E-M1] Eardley, D., Moncrief, V., The global existence of Yang-Mills-Higgs fields in 4-dimensional Minkowski space. I. Local existence and smoothness properties. *Comm. Math. Phys.* **83** (2) (1982), 171–191.

[E-M2] Eardley, D., Moncrief, V., The global existence of Yang-Mills-Higgs fields in 4-dimensional Minkowski space. II. Completion of proof. *Comm. Math. Phys.* **83** (2) (1982), 193–212.

[FKSY] Finster, F., Kamran, N., Smoller, J., Yau, S. T., Decay of solutions of the wave equation in Kerr geometry. Preprint.

[F-M] Fischer, A., Marsden, J., The Einstein evolution equations as a first-order quasi-linear symmetric hyperbolic system, I. *Comm. Math. Phys.* **28** (1972), 1–38.

[F] H. G. Friedlander *The Wave Equation on a Curved Space-time*. Cambridge Monogr. Math. Phys. 2, Cambridge University Press, Cambridge, New York, Melbourne 1975.

[Fr1] Friedrich, H., On the existence of n-geodesically complete or future complete solutions of Einstein's field equations with smooth asymptotic structure. *Comm. Math. Phys.* **107** (1986), 587–609.

[Fr2] Friedrich, H., Existence and structure of past asymptotically simple solutions of Einstein's field equations with positive cosmological constant. *J. Geom. Phys.* **3** (1986), 101–117 .

[H] Hörmander, L., The lifespan of classical solutions of nonlinear hyperbolic equations. In *Pseudodifferential operators* (Oberwolfach, 1986), Lecture Notes in Math. 1256, Springer-Verlag, Berlin 1987, 214–280.

[H-K-M] Hughes, T. J. R., Kato, T., and Marsden, J. E., Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity. *Arch. Rational Mech. Anal.* **63** (1977), 273–394.

[J] John, F., Blow-up of solutions of nonlinear wave equations in three space dimensions. *Manuscripta Math.* **28** (1979), 235–265.

[K-W] Kay, B., and Wald, R., Linear stability of Schwarzschild under perturbations which are nonvanishing on the bifurcation 2-sphere. *Classical Quantum Gravity* **4** (4) (1987), 893–898.

[K1] Klainerman, S., Long time behavior of solutions to nonlinear wave equations. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 2, PWN, Warsaw 1984, 1209–1215.

[K2] Klainerman, S., The null condition and global existence to nonlinear wave equations. *Lectures in Appl. Math.* **23** (1986), 293–326.

[K3] Klainerman, S., PDE as a unified subject. *Geom. Funct. Anal.* (2000), Special Volume, 279–315.

[K4] Klainerman, S., A commuting vectorfield approach to Strichartz type inequalities and applications to quasilinear wave equations. *Internat Math. Res. Notices* **2001** (5) (2001), 221–274.

[K-N1] Klainerman, S., and Nicolo, F., *The evolution problem in general relativity.* Progr. Math. Phys. 25, Birkhäuser, Boston, MA, 2003.

[K-N2] Klainerman, S., and Nicolo, F., Peeling properties of asymptotically flat solutions to the Einstein vacuum equations. *Classical Quantum Gravity*, to appear.

[K-R1] Klainerman, S., and Rodnianski, I., Improved local well posedness for quasilinear wave equations in dimension three. *Duke Math. J.* **117** (2003), 1–124.

[K-R2] Klainerman, S., and Rodnianski, I., Causal geometry of Einstein-Vacuum spacetimes with finite curvature flux. *Invent. Math.* **159** (2005), 437–529.

[K-R3] Klainerman, S., and Rodnianski, I., A geometric approach to Littlewood-Paley theory. *Geom. Funct. Anal.* **16** (1) (2006), 126–163.

[K-R4] Klainerman, S., and Rodnianski, I., Sharp trace theorems for null hypersurfaces on Einstein metrics with finite curvature flux. *Geom. Funct. Anal.* **16** (1) (2006), 164–229.

[K-R5] Klainerman, S., and Rodnianski, I., Lower bounds for the radius of injectivity of null hypersurfaces. Preprint.

[K-R6] Klainerman, S., and Rodnianski, I., A Kirchoff-Sobolev parametrix for the wave equations in a curved space-time. Preprint.

[K-R7] Klainerman, S., and Rodnianski, I., On the large data break-down criterion in General Relativity. In preparation.

[L1] Lindblad, H., Global solutions of nonlinear wave equations. *Comm. Pure Appl. Math.* **45** (9) (1992), 1063–1096.

[L2] Lindblad, H., Global solutions of quasilinear wave equations. Preprint.

[L-R1] Lindblad, H., and Rodnianski, I., The weak null condition for Einstein's equations. *C. R. Math. Acad. Sci. Paris* **336** (11) (2003), 901–906.

[L-R2] Lindblad, H., and Rodnianski, I., The global stability of Minkowski space-time in harmonic gauge. *Ann. of Math.*, to appear.

[M-S] Machedon, M., and Stalker, J., Decay of solutions to the wave equation on a spherically symmetric background. Preprint.

[Ma] Maxwell, D., Rough solutions of the Einstein constraint equations. *J. Reine Angew. Math.* **590** (2006), 1–29.

[Mo] Moncrief, V., An integral equation for spacetime curvature in General Relativity. Preprint.

[P-S] Ponce, G., and Sideris, T., Local regularity of nonlinear wave equations in three space dimensions. *Comm. Partial Differential Equations* **18** (1993), 169–177.

[Pr] Price, R., Nonspherical perturbations of relativistic gravitational collapse. I. Scalar and gravitational perturbations. *Phys. Rev. D* (3) **5** (1972), 2419–2438.

[Re] Reiris, M., Aspects of the long time evolution in General Relativity and geometrization of three-manifolds. PhD. Thesis, Stony Brook, 2005.

[Ri] Ringström, H., Strong cosmic censorship in $T^3$-Gowdy spacetimes. *Ann. of Math.*, to appear.

[S-Y1] Schoen, R., and Yau, S. T., On the proof of the positive mass conjecture in general relativity. *Comm. Math. Phys.* **65** (1979), 45–76.

[S-Y2] Schoen, R., and Yau, S. T., The existence of a black hole due to condensation of matter. *Comm. Math. Phys.* **90** (1983), 575–579.

[Sm] Smith, H., A parametrix construction for wave equations with $C^{1,1}$ coefficients. *Ann. Inst. Fourier* (*Grenoble*) **48** (3) (1998), 797–835.

[S-T1] Smith, H., and Tataru, D., Sharp counterexamples for Strichartz estimates for low regularity metrics. *Math. Res. Lett.* **9** (2002), 199–2004.

[S-T2] Smith, H., and Tataru, D., Sharp local well-posedness results for the nonlinear wave equation. *Ann. of Math.* **162** (2005), 291–366.

[Sob] Sobolev, S., Méthode nouvelle à resoudre le problème de Cauchy pour les équations linéaires hyperboliques normales. *Mat. Sb.* **1** (43) (1936), 31–79.

[Ta2] Tataru, D., Strichartz estimates for second order hyperbolic operators with non smooth coefficients. *Amer. J. Math.* **123** (2001), 385–423.

[Ta1] Tataru, D., Strichartz estimates for operators with non smooth coefficients and the nonlinear wave equation. *Amer. J. Math.* **122** (2000), 349–376.

[Tw] Twainy, F., The Time Decay of Solutions to the Scalar Wave Equation in Schwarzschild Background. Thesis, University of California, San Diego, 1989.

[W] Witten, E., A new proof of the positive mass theorem. *Comm. Math. Phys.* **80** (1981), 381–402.

[Z] N. Zipser, The Global Nonlinear Stability of the Trivial Solution of the Einstein-Maxwell Equations. PhD Thesis, Harvard University, 2000.

Department of Mathematics, Princeton University, Princeton, NJ 08544, U.S.A.
E-mail: irod@math.princeton.edu

# Categorification and correlation functions in conformal field theory

Christoph Schweigert, Jürgen Fuchs, and Ingo Runkel*

**Abstract.** A modular tensor category provides the appropriate data for the construction of a three-dimensional topological field theory. We describe the following analogue for two-dimensional conformal field theories: a 2-category whose objects are symmetric special Frobenius algebras in a modular tensor category and whose morphisms are categories of bimodules. This 2-category provides sufficient ingredients for constructing all correlation functions of a two-dimensional rational conformal field theory. The bimodules have the physical interpretation of chiral data, boundary conditions, and topological defect lines of this theory.

**Mathematics Subject Classification (2000).** Primary 81T40; Secondary 18D10, 18D35, 81T45.

**Keywords.** (Rational) conformal field theory, topological field theory, tensor categories, categorification.

## 1. Quantum field theories as functors

In approaches to quantum field theory that are based on the concepts of fields and states, the utility of categories and functors is by now well-established. The following pattern has been recognized: There is a geometric category $\mathcal{G}$ which, for every concrete model, must be suitably "decorated". The decoration is achieved with the help of objects and morphisms from another category $\mathcal{C}$. For known classes of quantum field theories, the decoration category $\mathcal{C}$ typically has a representation-theoretic origin – the reader is encouraged to think of it as the representation category of some algebraic object, like a quantum group, a loop group, a vertex algebra, a net of observable algebras, etc. This way one obtains a decorated geometric category $\mathcal{G}_\mathcal{C}$. The quantum field theory can then be formulated as a (tensor) functor $qft_\mathcal{C}$ from $\mathcal{G}_\mathcal{C}$ to some category of vector spaces. In this contribution, we mainly consider cases for which this latter category is the tensor category of finite-dimensional complex vector spaces.

A prototypical example for this pattern is provided by topological quantum field theories (TFTs). For such theories, the geometric category $\mathcal{G}$ is based on a cobordism category: its objects are $d-1$-dimensional topological manifolds without boundary. It is convenient to include two types of morphisms [42]: homeomorphisms of $d-1$-dimensional manifolds, and cobordisms. A cobordism $\mathrm{M}\colon \mathrm{Y}_1 \to \mathrm{Y}_2$ is a $d$-dimen-

*C. S. is supported by the DFG project SCHW 1162/1-1, and J. F. by VR under project no. 621–2003–2385.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

sional topological manifold M together with a parametrization of its boundary given by a homeomorphism $\partial M \xrightarrow{\cong} \overline{Y}_1 \sqcup Y_2$, where $\overline{Y}_1$ has the orientation opposite to the one of $Y_1$. The composition of morphisms is by concatenation, by gluing, and by changing the parametrization of the boundary, respectively. Cobordisms that coincide by a homeomorphism of the $d$-manifold M restricting to the identity on $\partial M$ must be identified.

In the simplest case, a topological field theory thus associates to a closed $d-1$-dimensional manifold X a vector space $qft_{\mathcal{C}}(X)$, and to a homeomorphism or a cobordism $M\colon Y_1 \to Y_2$ a linear map

$$qft_{\mathcal{C}}(M)\colon qft_{\mathcal{C}}(Y_1) \to qft_{\mathcal{C}}(Y_2).$$

The assignment $qft_{\mathcal{C}}$ is required to be a (strict) tensor functor. This requirement implies the usual axioms (cf. e.g. [43]) of naturality, multiplicativity, functoriality and normalization.

There also exists a path-integral approach to certain classes of topological field theories. Its relation to the categorical framework described above is as follows: One can think of the vector space $qft_{\mathcal{C}}(\partial_- M)$ as the space of (equivalence classes of) possible initial data for "fields" in the path integral, and of $qft_{\mathcal{C}}(\partial_+ M)$ as the possible final data. The matrix elements of the linear map $qft_{\mathcal{C}}(M)$ are then the transition amplitudes for fixed initial and final values of the fields.

This picture is still oversimplified. In particular, it turns out that it is natural to enrich also the geometric category $\mathcal{G}$ over the category of complex vector spaces. As a consequence, when studying functors on $\mathcal{G}$, one should then consider also projective functors. These issues, which are closely related to anomalies in quantum field theory will, however, be suppressed in this note.

A prominent class of examples of 3-dimensional topological field theories arises from Chern–Simons field theories. For $G$ a simple connected and simply-connected complex Lie group, consider holomorphic $G$-bundles on a closed two-manifold X of genus $g$ with complex structure. Pick a generator $\mathcal{L}$ for the Picard group of the moduli space $\mathcal{M}_X^G$ of such bundles. Upon changing the complex structure of X, the vector spaces $H^0(\mathcal{M}_X^G, \mathcal{L}^{\otimes k})$ fit together into a vector bundle with projectively flat connection over the moduli space $\mathcal{M}_g^G$ of curves of genus $g$. The complex modular functor [3] associates these bundles to X; these bundles and their monodromies provide a formalization of all aspects of the chiral level-$k$ Wess–Zumino–Witten (WZW) conformal field theory for $G$ that are needed for the discussions in the subsequent sections.

As a next step, it is natural to extend the formalism by allowing for marked points with additional structure on the two-manifolds. From a field theoretical point of view, this is motivated by the desire to account for insertions of fields. In the case of Chern–Simons theory, the additional structure amounts to specifying parabolic structures at the marked points. The marked points thus have to carry labels, which we will identify in a moment as objects of a decoration category $\mathcal{C}$.

This structure must be extended to the geometric morphisms: Maps of 2-dimensional manifolds are required to preserve marked points and the decoration in $\mathcal{C}$. The decoration of the 2-dimensional manifolds is extended to the 3-dimensional manifolds M underlying cobordisms by supplying them with oriented (ribbon) graphs ending on (arcs through) the marked points on $\partial_{\pm}$M. The ribbon graph is allowed to have vertices with a finite number of ingoing and outgoing ribbons. From the construction of invariants of knots and links, it is known that this enforces a categorification of the set of labels: $\mathcal{C}$ must be a ribbon category, i.e. a braided sovereign tensor category. In particular, the vertices of the graph are to be labeled by morphisms in the decoration category $\mathcal{C}$.

This approach has been very fruitful and has, in particular, made a rigorous construction of Chern–Simons theory possible [38], [43]. The extension from invariants of links in $\mathbb{R}^3$ to link invariants in arbitrary oriented three-manifolds has revealed an important subclass of tensor categories: modular tensor categories.

For the purposes of the present contribution, we adopt the following definition of a modular tensor category: it is an abelian, $\mathbb{C}$-linear, semi-simple ribbon category with a finite number of isomorphism classes of simple objects. The tensor unit $\mathbf{1}$ is required to be simple, and the braiding must be nondegenerate in the sense that the natural transformations of the identity functor on $\mathcal{C}$ are controlled by the fusion ring $K_0(\mathcal{C})$:

$$\mathrm{End}(\mathrm{Id}_{\mathcal{C}}) \cong K_0(\mathcal{C}) \otimes_{\mathbb{Z}} \mathbb{C}.$$

The relation between Chern–Simons theory and chiral Wess–Zumino–Witten theory [44] was a first indication that modular tensor categories also constitute the appropriate mathematical formalization of the chiral data [31], [16] of a conformal field theory. Recent progress in representation theory has made this idea much more precise; for the following classes of representation categories it has been established that they carry the structure of a modular tensor category:

- The representation category of a connected ribbon factorizable weak Hopf algebra over $\mathbb{C}$ (or, more generally, over an algebraically closed field $\Bbbk$) with a Haar integral [33].
- The category of unitary representations of the double of a connected $C^*$ weak Hopf algebra [33].
- The category of local sectors of a finite-index net of von Neumann algebras on the real line, if the net is strongly additive (which for conformal nets is equivalent to Haag duality) and has the split property [27].
  (In this example and in the previous one one obtains unitary modular tensor categories.)
- The representation category of a self-dual vertex algebra that obeys Zhu's $C_2$ cofiniteness condition and certain conditions on its homogeneous subspaces, provided that this category is semisimple [25].

The last two entries in this list correspond to two different mathematical formalizations of chiral conformal field theories. The results of [27] and [25] therefore justify the point of view that modular tensor categories constitute an œcumenic formalization of the chiral data of a conformal field theory.

## 2. Two-dimensional conformal field theories

Three-dimensional topological field theory will indeed appear as a tool in constructions below. Our main interest here is, however, in a different class of quantum field theories: full (and in particular local) two-dimensional conformal field theories, or CFTs, for short.

For these theories, the geometric category of interest is the category of two-dimensional conformal manifolds, possibly with non-empty boundary. This fact already indicates that full conformal field theories are different from the chiral conformal field theories that we encountered in the last section for the case of WZW theories. Morphisms in the category of conformal manifolds are maps respecting the conformal structure. Actually, there are two different types of full conformal field theories, corresponding to two different geometric categories: One considers either oriented conformal manifolds, leading to a category $\mathcal{G}^{\mathrm{or}}$, or unoriented manifolds, leading to a different geometric category $\mathcal{G}^{\mathrm{unor}}$. As morphisms, we admit maps that preserve the respective structure. (In the application of conformal field theory to string theory, $\mathcal{G}^{\mathrm{or}}$ plays a role in superstrings of "type II", while $\mathcal{G}^{\mathrm{unor}}$ appears in superstring theories of "type I".)

As in the case of topological field theories, the geometric category needs to be decorated. To find the appropriate decoration data, we first discuss a physical structure that is known to be present in specific classes of models and that our approach to conformal field theory should take into account:

- Whenever a two-manifold X has a boundary, one expects that it is necessary to specify boundary conditions. In a path integral approach, a boundary condition is a prescription for the boundary values of fields that appear in the Lagrangian. Here, a more abstract approach is adequate: we take the possible boundary conditions to be the objects of a decoration category $\mathcal{M}$. This constitutes again a categorification of the decoration data. It can be motivated further by the observation that insertions of boundary fields can change the boundary condition; they will be related to morphisms of the category $\mathcal{M}$.

A second structure, which in the literature has received much less attention than boundary conditions, turns out to provide crucial clues for the construction of full conformal field theories:

- Conformal field theories can have topological defect lines. Such defect lines have e.g. been known for the Ising model for a long time: This CFT describes

the continuum limit of a lattice model with $\mathbb{Z}_2$-valued variables at the vertices of a two-dimensional lattice and with ferromagnetic couplings along its edges. A defect line is obtained when one changes the couplings on all edges that intersect a given line in the lattice from ferromagnetic to antiferromagnetic.

In the continuum limit, such a defect line can be described by a condition on the values of bulk fields at the defect line. In particular, when crossing a defect line, the correlation function of a bulk field can acquire a branch cut. Indeed, the defect lines we are interested in behave very much like branch cuts: they are topological in the sense that their precise location does not matter. In a field theoretic language, this is attributed to the fact that the stress-energy tensor of the theory is required to be smooth across defect lines. As in the case of boundary conditions, in our framework it is not desirable to express defect lines through conditions on the values of fields. Instead, we anticipate again a categorification of the decoration data and label the possible types of defect lines by objects in yet another decoration category $\mathcal{D}$.

There is a natural notion of fusion of defect lines, see e.g. [36], [15]. Accordingly, $\mathcal{D}$ will be a tensor category. Also, to take into account the topological nature of defect lines, we assume that the tensor category $\mathcal{D}$ has dualities and that it is even sovereign. In contrast, there is no natural notion of a braiding of defect lines, so $\mathcal{D}$ is, in general, not a ribbon category.

The two decoration categories – $\mathcal{D}$ for the defects and $\mathcal{M}$ for the boundary conditions – are themselves related. One can fuse a defect line to a boundary condition, thereby obtaining another boundary condition; see e.g. [24]. This endows the category $\mathcal{M}$ of boundary conditions with the structure of a module category over $\mathcal{D}$, i.e. there is a bifunctor

$$\otimes\colon \mathcal{D} \times \mathcal{M} \to \mathcal{M}$$

which has (mixed) associativity constraints obeying a mixed pentagon identity.

The structure just unraveled – a tensor category $\mathcal{D}$ together with a module category $\mathcal{M}$ over $\mathcal{D}$ – calls for the following natural extension: One should also consider the category of module functors, i.e. the category $\mathcal{C}$ whose objects are endofunctors of $\mathcal{M}$ that are compatible with the structure of a module category over $\mathcal{D}$ and whose morphisms are natural transformations between such functors. The concatenation of functors naturally endows $\mathcal{C}$ with a product so that $\mathcal{C}$ is a tensor category.

A recent insight is the following: In the application to two-dimensional conformal field theory, the category $\mathcal{C}$ obtained this way is equivalent to the category of chiral data that we discussed in Section 1! There is a particularly amenable subclass of conformal field theories, called *rational* conformal field theories (*R*CFTs), which can be rigorously discussed on the basis of this idea. For these theories, the category $\mathcal{C}$ of chiral data has the structure of a modular tensor category. In this case the idea can be exploited to arrive at a construction of correlation functions (see Section 4) of rational conformal field theories that is based on three-dimensional topological field theory. This TFT approach to RCFT correlators will be presented in Section 5 below.

## 3. The 2-category of Frobenius algebras

In practice, one frequently takes an opposite point of view: Instead of obtaining $\mathcal{C}$ as a functor category, one starts from some knowledge about the chiral symmetries of a conformal field theory. This allows one to use the representation-theoretic results mentioned in Section 1 so as to get a modular tensor category $\mathcal{C}$ describing the chiral data of the theory. Afterwards one realizes that the category $\mathcal{M}$ of boundary conditions is also a (right-) module category over $\mathcal{C}$. General arguments involving internal Hom's [35] together with specific properties relevant to rational conformal field theories then imply that in the tensor category $\mathcal{C}$ there exists an associative algebra $A$ such that $\mathcal{M}$ is equivalent to the category $\mathcal{C}_A$ of left $A$-modules in $\mathcal{C}$. By similar arguments one concludes that the category $\mathcal{D}$ is equivalent to the category of $A$-bimodules. Additional constraints, in particular the non-degeneracy of the two-point functions of boundary fields on a disk, lead to further conditions on this algebra [17]: $A$ must be a symmetric special Frobenius algebra. Owing to these insights we are able to use a generalization[1] of the theory of Frobenius algebras to braided tensor categories as a powerful tool to analyze (rational) conformal field theory. The algebraic theory in the braided setting is, however, genuinely richer; see [14] for a discussion of some new phenomena.

A *Frobenius* algebra $A = (A, m, \eta, \Delta, \varepsilon)$ in $\mathcal{C}$ is, by definition, an object of $\mathcal{C}$ carrying the structures of a unital associative algebra $(A, m, \eta)$ and of a counital coassociative coalgebra $(A, \Delta, \varepsilon)$ in $\mathcal{C}$, with the algebra and coalgebra structures satisfying the compatibility requirement that the coproduct $\Delta\colon A \to A \otimes A$ is a morphism of $A$-bimodules (or, equivalently, that the product $m\colon A \otimes A \to A$ is a morphism of $A$-bi-comodules). A Frobenius algebra is called *special* iff the coproduct is a right-inverse to the product – this means in particular that the algebra is separable – and a nonvanishing multiple of the unit $\eta\colon \mathbf{1} \to A$ is a right-inverse to the counit $\varepsilon\colon A \to \mathbf{1}$. There are two isomorphisms $A \to A^\vee$ that are naturally induced by product, counit and duality; $A$ is called *symmetric* iff these two isomorphisms coincide.

Two algebras in a tensor category $\mathcal{C}$ are called Morita equivalent iff their representation categories are equivalent as module categories over $\mathcal{C}$. Since the algebra $A$ is characterized by the requirement that $\mathcal{C}_A$ is equivalent to the given decoration category $\mathcal{M}$, it is clear that only the Morita class of the algebra should matter. It is a non-trivial internal consistency check on the constructions to be presented in Section 5 that this is indeed the case.

Taking the modular tensor category $\mathcal{C}$ as the starting point, the following further generalization of the setup (compare also [32], [46], [28]) is now natural:[2] Consider the set of *all* (symmetric special) Frobenius algebras in $\mathcal{C}$. This gives rise to a *family* of full conformal field theories that are based on the same chiral data. And

[1] When $\mathcal{C}$ is the modular tensor category of finite-dimensional complex vector spaces, the CFT is a topological CFT. In particular, $A$ is then an ordinary Frobenius algebra. This case has served as a toy model for conformal field theories, see e.g. [40], [30].

[2] We are grateful to Urs Schreiber for discussions on this point.

again we categorify the structure: we introduce a 2-category $\mathcal{F}rob_{\mathcal{C}}$ whose objects are symmetric special Frobenius algebras in $\mathcal{C}$. The 1-morphisms $\mathcal{H}om(A, A')$ are given by the category of $A$-$A'$-bimodules. The 2-category $\mathcal{F}rob_{\mathcal{C}}$ has a distinguished object $I$: as an object of $\mathcal{C}$, $I$ is just the tensor unit, which is naturally a symmetric special Frobenius algebra. Because of the considerations in [8], the full conformal field theory corresponding to $I$ is often referred to as the "Cardy case"; for this case a construction of the correlators in the spirit of Section 4 was given in [11].

We are now in a position to attribute a physical interpretation to the morphisms of $\mathcal{F}rob_{\mathcal{C}}$. $\mathcal{H}om(I, I)$ is naturally identified with the tensor category $\mathcal{C}$ of chiral data. Further, for any $A$ the tensor category $\mathcal{H}om(A, A)$ describes topological defects in the full conformal field theory associated to $A$; more generally, the category $\mathcal{H}om(A, A')$ accounts for topological defect lines that separate two different conformal field theories which share the same chiral data. Finally, the category $\mathcal{H}om(I, A)$ also describes boundary conditions for the full conformal field theory labeled by $A$.

We have thus learned that the decoration data of a family of full rational conformal field theories based on the same chiral data are described by a 2-category. This nicely fits with insight gained in other contexts:

- 2-categories appear in recent approaches to elliptic objects [2], [42].
- Hermitian bundle gerbes, which appear naturally in a semi-classical description of WZW conformal field theories [23], form a 2-category [41].

Unfortunately, at the time of writing, a unified approach to conformal field theories based on 2-categories has not been established yet. For this reason, in the sequel we will not be able to use this language systematically.

We close this section with two further comments. First, so far we have discussed the decoration data relevant to the oriented geometric category $\mathcal{G}^{\mathrm{or}}$. For the unoriented geometric category $\mathcal{G}^{\mathrm{unor}}$, additional structure on the relevant Frobenius algebra is needed: $A$ must then be a *Jandl* algebra, that is, a symmetric special Frobenius algebra coming with an algebra isomorphism $A \xrightarrow{\cong} A^{\mathrm{opp}}$ that squares to the twist on $A$. This turns out to be the appropriate generalization of the notion of an algebra with involution to braided tensor categories. We refrain from discussing this issue in the present contribution, but rather refer to [18], [21] for details.

Second, the general situation encountered above – a module category $\mathcal{M}$ over a tensor category $\mathcal{C}$ – naturally appears in various other contexts as well:

- The left modules over a weak Hopf algebra $H$ form a tensor category $\mathcal{C} = H$-Mod. In a weak Hopf algebra, one can identify two subalgebras $H_s$ and $H_t$ that are each other's opposed algebras [6]. Forgetful functors thus endow any $H$-module with the structure of an $H_t$-bimodule; one even obtains a tensor functor from $H$-Mod to $H_t$-Bimod. The usual tensor product of a $H_t$-bimodule and an $H_t$-left module endows the category of $H_t$-modules with the structure of a module category over $H_t$-Bimod and thus over $H$-Mod.

Weak Hopf algebras have indeed been proposed [37] as a framework for rational conformal field theories. Unfortunately, such a description must cope with two problems: First, to account for a braiding on $\mathcal{C}$, one must work with an $R$-matrix on $H$; not too surprisingly, this is technically involved, and indeed not very much is known about $R$-matrices on weak Hopf algebras. Secondly, given a tensor category $\mathcal{C}$ (describing the chiral data), there does not exist a canonical weak Hopf algebra such that $H$-Mod is equivalent to $\mathcal{C}$. Rather, as an additional datum, a fiber functor to $H_t$-bimodules needs to be chosen. A physical interpretation of this datum is unclear. On the other hand, Hopf algebras are still useful in the analysis of full rational CFT: Their Hochschild cohomology was used [10] to compute the Davydov–Yetter cohomology of the pair $(\mathcal{C}, \mathcal{M})$; from the vanishing of this cohomology, rigidity properties of rational conformal field theories follow.

- Weak Hopf algebras also appear in the study of inclusions of subfactors. For a review and further references, we refer to Sections 8 and 9 of [34].
- The same category-theoretic structures have been recovered in the theory of vertex algebras from so-called open-string vertex algebras [26] which are, in particular, extensions of ordinary vertex algebras.

Not surprisingly, some of the structures that will be encountered in the rest of this paper also have counterparts in the context of weak Hopf algebras, of nets of subfactors, or of open-string vertex algebras.

## 4. Correlation functions

The observations made in the preceding section raise the question whether one can construct a full rational conformal field theory by using a modular tensor category $\mathcal{C}$ and the 2-category $\mathcal{F}rob_{\mathcal{C}}$ as an input. These data should then in particular encode information about the correlation functions of the conformal field theory.

To decide this question, it is helpful to reformulate first the geometric categories $\mathcal{G}^{\mathrm{or}}$ and $\mathcal{G}^{\mathrm{unor}}$. This is achieved with the help of a crucial aspect of complex geometry in *two* dimensions: a complex structure on a two-dimensional manifold is equivalent to a conformal structure and the choice of an orientation. The complex double $\widehat{\mathrm{X}}$ of a conformal manifold X is a two-sheeted cover of X whose points are pairs consisting of a point $p \in \mathrm{X}$ and a local orientation at $p$. In view of the previous comment, it is clear that $\widehat{\mathrm{X}}$ is a complex curve. We have thus associated to any object X of the geometric categories $\mathcal{G}^{\mathrm{unor}}$ and $\mathcal{G}^{\mathrm{or}}$ a complex curve $\widehat{\mathrm{X}}$ that comes with an orientation reversing involution $\sigma$ such that the quotient $\widehat{\mathrm{X}}/\langle\sigma\rangle$ is naturally isomorphic to X, and we have a canonical projection

$$\pi : \widehat{\mathrm{X}} \mapsto \mathrm{X} \cong \widehat{\mathrm{X}}/\langle\sigma\rangle.$$

The set of fixed points of $\sigma$ is just the preimage under $\pi$ of the boundary $\partial\mathrm{X}$.

To be able to use the tools of complex geometry, we therefore reformulate our geometric categories as follows: in the case of $\mathcal{G}^{\mathrm{unor}}$, the objects are pairs $(\widehat{\mathrm{X}}, \sigma)$ consisting of a complex curve $\widehat{\mathrm{X}}$ and an anticonformal involution $\sigma$ that implements the action of the Galois group of $\mathbb{C}/\mathbb{R}$. In the case of $\mathcal{G}^{\mathrm{or}}$, we fix a global section of $\pi$ as an additional datum.

Next, since we are interested in correlation functions depending on insertion points, it is natural to consider simultaneously the family $\mathcal{M}_{g,m}$ of all complex curves with marked points that have the same topological type (i.e., genus $g$ and number $m$ of marked points) as $\widehat{\mathrm{X}}$. It is convenient to treat the positions of the insertion points and the moduli of the complex structure on the same footing. The curves that admit an involution $\sigma$ of the same type as $\widehat{\mathrm{X}}$ and for which the marked points are related by $\sigma$ form a submanifold $\mathcal{M}^{\sigma}_{g,m}$ of $\mathcal{M}_{g,m}$. (To be precise, one obtains [7] such a relation for Teichmüller spaces, rather than for moduli spaces.)

Given the modular tensor category $\mathcal{C}$, the complex modular functor [3] provides us with a vector bundle $\mathcal{V}$ with projectively flat connection on $\mathcal{M}_{g,m}$. We can now formulate the 'principle of holomorphic factorization' (which for certain classes of conformal field theories follows from chiral Ward identities that can formally be derived from an action functional [45]). It states that, first of all, the conformal surface X should be decorated in such a way that the double $\widehat{\mathrm{X}}$ has the structure of an object in the decorated cobordism category for the topological field theory based on $\mathcal{C}$. It then makes sense to require, secondly, that the correlation function is a certain global section of the restriction of $\mathcal{V}$ to $\mathcal{M}^{\sigma}_{g,m}$.

At this point, it proves to be convenient to use the equivalence of the complex modular functor and the topological modular functor $tft_{\mathcal{C}}$ based on the modular tensor category $\mathcal{C}$ [3] so as to work in a topological (rather than complex-analytic) category. We are thereby lead to the description of a correlation function on X as a specific vector $\mathrm{Cor}(\mathrm{X})$ in the vector space $tft_{\mathcal{C}}(\widehat{\mathrm{X}})$ that is assigned to the double $\widehat{\mathrm{X}}$ by the topological modular functor $tft_{\mathcal{C}}$. These vectors must obey two additional axioms:

- *Covariance*: Given any morphism $f : \mathrm{X} \to \mathrm{Y}$ in the relevant decorated geometric category $\mathcal{G}_{\mathcal{C}}$, we demand

$$\mathrm{Cor}(\mathrm{Y}) = tft_{\mathcal{C}}(f)(\mathrm{Cor}(\mathrm{X})).$$

- *Factorization*: Certain factorization properties must be fulfilled.

  We refer to [12], [13] for a precise formulation of these constraints.

The covariance axiom implies in particular that the vector $\mathrm{Cor}(\mathrm{X})$ is invariant under the action of the mapping class group $\mathrm{Map}(\mathrm{X}) \cong \mathrm{Map}(\widehat{\mathrm{X}})^{\sigma}$. This group, also called the relative modular group [4], acts genuinely on $tft_{\mathcal{C}}(\widehat{\mathrm{X}})$.

## 5. Surface holonomy

To find solutions to the covariance and factorization constraints on the vectors $\mathrm{Cor}(\mathrm{X}) \in tft_{\mathcal{C}}(\widehat{\mathrm{X}})$ we use the three-dimensional topological field theory associated to the modular tensor category $\mathcal{C}$. Thus we look for a (decorated) cobordism $(\mathrm{M_X}, \emptyset, \widehat{\mathrm{X}})$ such that the vector $tft_{\mathcal{C}}(\mathrm{M_X}, \emptyset, \widehat{\mathrm{X}})1 \in tft_{\mathcal{C}}(\widehat{\mathrm{X}})$ is the correlator $\mathrm{Cor}(\mathrm{X})$.

The three-manifold $\mathrm{M_X}$ should better not introduce any topological information that is not already contained in X. This leads to the idea to use an interval bundle as a "fattening" of the world sheet. It turns out that the following quotient of the interval bundle on $\widehat{\mathrm{X}}$, called the connecting (three-) manifold, is appropriate [11]:

$$\mathrm{M_X} = (\,\widehat{\mathrm{X}} \times [-1, 1]\,)/\langle(\sigma, t \mapsto -t)\rangle.$$

This three-manifold is oriented, has boundary $\partial \mathrm{M_X} \cong \widehat{\mathrm{X}}$, and it contains X as a retract: the embedding $\iota$ of X is to the fiber $t = 0$, the retracting map contracts along the intervals.

The connecting manifold $\mathrm{M_X}$ must now be decorated with the help of the decoration categories $\mathcal{H}om(A, A')$. We will describe this procedure for the oriented case only. The conformal surface X is decomposed by defect lines (which are allowed to end on $\partial$X) into various two-dimensional regions. There are two types of one-dimensional structures: boundary components of X and defect lines. Defect lines, in general, form a network; they can be closed or have end points, and in the latter case they can end either on the boundary or in the interior of X. Both one-dimensional structures are partitioned into segments by marked "insertion" points. The end points of defect lines carry insertions, too. Finally, we also allow for insertion points in the interior of two-dimensional regions.

To these geometric structures, data of the 2-category $\mathcal{F}rob_{\mathcal{C}}$ are now assigned as follows. First, we attach to each two-dimensional region a symmetric special Frobenius algebra, i.e. an object of $\mathcal{F}rob_{\mathcal{C}}$. To a segment of a defect line that separates regions with label $A$ and $A'$, respectively, we associate a 1-morphism in $\mathcal{H}om(A, A')$, i.e. an $A$-$A'$-bimodule. Similarly, to a boundary segment adjacent to a region labeled by $A$, we assign an object in $\mathcal{H}om(I, A)$, i.e. a left $A$-module. Finally, zero-dimensional geometric objects are labeled with 2-morphisms of $\mathcal{F}rob_{\mathcal{C}}$; in particular, junctions of defect lines with each other or with a boundary segment are labeled by 2-morphisms from the obvious spaces.

Two types of points, however, still deserve more comments: those separating boundary segments on the one hand, and those separating or creating segments of defect lines or appearing in the interior of two-dimensional regions on the other. These are the *insertion points* that were mentioned above. An insertion point $p \in \partial$X that separates two boundary segments labeled by objects $M_1, M_2 \in \mathcal{H}om(I, A)$ has a single preimage under the canonical projection $\pi$ from $\widehat{\mathrm{X}}$ to X; to the interval in $\mathrm{M_X}$ that joins this preimage to the image $\iota(p)$ of $p$ under the embedding $\iota$ of X into $\mathrm{M_X}$, we assign an object $U$ of the category $\mathcal{C} = \mathcal{H}om(I, I)$ of chiral data. To the insertion

point itself, we then attach a 2-morphism in the morphism space $\mathrm{Hom}(M_1 \otimes U, M_2)$ in $\mathcal{H}om(I, A)$.

An insertion point in the interior of X has two preimages on $\widehat{\mathrm{X}}$; these two points are connected to $\iota(p)$ by two intervals. To each of these two intervals we assign to each of these two intervals an object $U$ and $V$, respectively, of the category $\mathcal{C}$ of chiral data. In the oriented case, the global section of $\pi$ is used to attribute the two objects $U$, $V$ to the two preimages. (For the unoriented case, the situation is more involved; in particular, the Jandl structure on the relevant Frobenius algebra enters the prescription.) We now first consider an insertion point separating a segment of a defect line labeled by an object $B_1 \in \mathcal{H}om(A, A')$ from a segment labeled by $B_2 \in \mathcal{H}om(A, A')$. We then use the left action $\rho_l$ of $A$ and the right action $\rho_r$ of $A'$ on the bimodule $B_1$ to define a bimodule structure on the object $U \otimes B_1 \otimes V$ of $\mathcal{C}$ by taking the morphisms $(\mathrm{id}_U \otimes \rho_l \otimes \mathrm{id}_V) \circ (c^{-1}_{U,A} \otimes \mathrm{id}_{B_1} \otimes \mathrm{id}_V)$ and $(\mathrm{id}_U \otimes \rho_r \otimes \mathrm{id}_V) \circ (\mathrm{id}_U \otimes \mathrm{id}_{B_1} \otimes c^{-1}_{A',V})$ as the action of $A$ and $A'$, respectively, where $c$ denotes the braiding isomorphisms of $\mathcal{C}$. The insertion point separating the defect lines is now labeled by a 2-morphism in $\mathrm{Hom}(U \otimes B_1 \otimes V, B_2)$, i.e. by a morphism of $A$-$A'$-bimodules.

To deal with insertion points in the interior of a two-dimensional region labeled by a Frobenius algebra $A$, we need to invoke one further idea: such a region has to be endowed with (the dual of) a triangulation $\Gamma$. To each edge of $\Gamma$ we attach the morphism $\Delta \circ \eta \in \mathrm{Hom}(\mathbf{1}, A \otimes A)$, and to each trivalent vertex of $\Gamma$ the morphism $\varepsilon \circ m \circ (m \otimes \mathrm{id}_A) \in \mathrm{Hom}(A \otimes A \otimes A), \mathbf{1})$. This pattern is characteristic for notions of surface holonomy. It has appeared in lattice topological field theories [22] and shows up in the surface holonomy of bundle gerbes as well. (For more details, references, and the relation to the Wess–Zumino term of WZW conformal field theories in a Lagrangian description, see [23].)

Now each of the insertion points $p$ that we still need to discuss is located inside a two-dimensional region labeled by some Frobenius algebra $A$ or creates a defect line. For the first type of points, we choose the triangulation such that an $A$-ribbon passes through $p$; to $p$ we then attach a bimodule morphism in $\mathrm{Hom}(U \otimes A \otimes V, A)$, with $U$ and $V$ objects of $\mathcal{C}$ as above. To a point $p$ at which a defect line of type $B$ starts or ends, we attach a bimodule morphism in $\mathrm{Hom}(U \otimes A \otimes V, B)$ and in $\mathrm{Hom}(U \otimes B \otimes V, A)$, respectively.

We have now obtained a complete labelling of a ribbon graph in the connecting manifold $\mathrm{M_X}$ with objects and morphisms of the modular tensor category $\mathcal{C}$; in other words, a cobordism from $\emptyset$ to $\widehat{\mathrm{X}}$ in the decorated geometric category $\mathcal{G}_{\mathcal{C}}$. Applying the modular functor for the tensor category $\mathcal{C}$ to this cobordism, we obtain a vector

$$\mathrm{Cor(X)} = tft_{\mathcal{C}}(\mathrm{M_X})\, 1 \in tft_{\mathcal{C}}(\widehat{\mathrm{X}}).$$

This is the prescription for RCFT correlation functions in the TFT approach. It follows from the defining properties of a symmetric special Frobenius algebra that Cor(X) does not depend on the choice of triangulation; for details see [12].

## 6. Results

On the basis of this construction one can establish many further results. Let us list some of them, without indicating any of their proofs:

[17] Of particular interest are the correlators for X being the torus or the annulus without field insertions, but possibly with defect lines. From these "one-loop amplitudes" one can derive concrete expressions for partition functions of boundary, bulk and defect fields.

The coefficients of these partition functions in the distinguished basis of the zero-point blocks on the torus that is given by characters can be shown to be equal to the dimensions of certain spaces of 2-morphisms of the 2-category $\mathcal{F}rob_{\mathcal{C}}$. Thus in particular they are non-negative integers.

In fact, one recovers expressions that had also been obtained in an approach based on subfactors [29], [5]. Moreover, these coefficients can be shown to satisfy other consistency requirements like forming so-called NIMreps of the fusion rules.

[18] To extend these results to unoriented (in particular, to unorientable) surfaces one must specify as additional datum a Jandl structure on the relevant Frobenius algebra. One can then e.g. compute the partition functions for the Möbius strip and Klein bottle. Their coefficients in the distinguished basis of zero-point torus blocks are integers, and for CFT models which serve as building blocks of type I string theories, these partition functions combine with the torus and annulus amplitudes in a way consistent with an interpretation in terms of state spaces of the string theory.

[19] The expressions for correlation functions can be made particularly explicit for conformal field theories of simple current type [39], which correspond to Frobenius algebras for which every simple subobject is invertible. Eilenberg–Mac Lane's [9] abelian group cohomology turns out to provide a crucial tool for analyzing this case.

It should be stressed, though, that the TFT approach to RCFT correlators treats the simple current case and other conformal field theories (i.e. those having an 'exceptional modular invariant' as their torus partition function) on an equal footing.

[20] By expressing some specific correlation functions for the sphere, the disk, and the real projective plane through the appropriate (two- or three-point) conformal blocks, one can derive explicit expressions for the coefficients of operator product expansions of bulk, boundary, and defect fields.

[12] For arbitrary topology of the surface X the correlators obtained in the TFT construction can be shown to satisfy the covariance and factorization axioms that were stated at the end of Section 4.

[15] The Picard group of the tensor category $\mathcal{H}om(A, A)$ describes symmetries of the full conformal field theory that is associated to $A$. The fusion ring $K_0(\mathcal{H}om(A, A))$ of that category contains information about Kramers–Wannier-like dualities as well.

## 7. Conclusions

The TFT approach to the construction of CFT correlation functions, which represents CFT quantities as invariants of knots and links in three-manifolds, relates a general paradigm of quantum field theory to the theory of (symmetric special) Frobenius algebras in (modular) tensor categories. It thereby constitutes a powerful algebraization of many questions that arise in the study of conformal field theory. As a result, one can both make rigorous statements about rational conformal field theories and set up efficient algorithms for the computation of observable CFT quantities.

A rich dictionary relating algebraic concepts and physical notions is emerging. It includes in particular the following entries:

- The classification of (oriented) full conformal field theories for given chiral data $\mathcal{C}$ amounts to the classification of Morita classes of Frobenius algebras in $\mathcal{C}$.
  As a special case, the classification of those theories whose torus partition function is "of automorphism type" amounts to determining the Brauer group of the category $\mathcal{C}$.

- The Picard group of the tensor category $\mathcal{H}om(A, A)$ acts as a symmetry group on the full conformal field theory associated to $A$, while the fusion ring of this tensor category contains information about Kramers–Wannier like dualities.

- Deformations of the conformal field theory are controlled by the Davydov–Yetter cohomology of the pair $(\mathcal{C}, \mathcal{C}_A)$.

The structure of this dictionary gives us confidence that some of the insights of the TFT approach – though, unfortunately, not most of the proofs – will still be relevant for the study of conformal field theories that are not rational any more.

## References

[1] Atiyah, M. F., Topological quantum field theories. *Inst. Hautes Études Sci. Publ. Math.* **68** (1988), 175–186.

[2] Baas, N. A., Dundas, B. I., and Rognes, J., Two-vector bundles and forms of elliptic cohomology. In *Topology, Geometry and Quantum Field Theory* (ed. by U. Tillmann), London Math. Soc. Lecture Note Ser. 308, Cambridge University Press, Cambridge 2004, 18–45.

[3] Bakalov B., and Kirillov, A. A., *Lectures on Tensor Categories and Modular Functors* Amer. Math. Soc., Providence, RI, 2001.

[4] Bianchi M., and Sagnotti, A., Open strings and the relative modular group. *Phys. Lett. B* **231** (1989), 389–396.

[5] Böckenhauer, J., Evans, D. E., and Kawahigashi, Y., Longo–Rehren subfactors arising from $\alpha$-induction. *Publ. Res. Inst. Math. Sci.* **37** (2001), 1–35.

[6] Böhm, G., Nill, F., and Szlachányi, K., Weak Hopf algebras: I. Integral theory and $C^*$-structure. *J. Algebra* **221** (1999), 385–438.

[7] Buser, P., and Seppälä, M., Real structures of Teichmüller spaces, Dehn twists, and moduli spaces of real curves. *Math. Z.* **232** (1999), 547–558.

[8] Cardy, J. L., Boundary conditions, fusion rules and the Verlinde formula. *Nucl. Phys. B* **324** (1989), 581–596.

[9] Eilenberg, S., and Mac Lane, S., Cohomology theory of abelian groups and homotopy theory II. *Proc. Natl. Acad. Sci. USA* **36** (1950), 657–663.

[10] Etingof, P. I., Nikshych, D., and Ostrik, V., On fusion categories. *Ann. of Math.* (2) **162** (2005), 581–642.

[11] Felder, G., Fröhlich, J., Fuchs, J., and Schweigert, C., Correlation functions and boundary conditions in RCFT and three-dimensional topology. *Compositio Math.* **131** (2002) 189–237.

[12] Fjelstad, J., Fuchs, J., Runkel, I., and Schweigert, C., TFT construction of RCFT correlators V: Proof of modular invariance and factorisation. Preprint; hep-th/0503194.

[13] —, Topological and conformal field theory as Frobenius algebras. Preprint; math. CT/0512076

[14] J. Fröhlich, Fuchs, J., I. Runkel, and C. Schweigert, Correspondences of ribbon categories. *Adv. Math.* **199** (2006), 192–329.

[15] —, Kramers–Wannier duality from conformal defects. *Phys. Rev. Lett.* **93** (2004), 070601.

[16] Fröhlich, J., and King, C., Two-dimensional conformal field theory and three-dimensional topology. *Int. J. Mod. Phys. A* **4** (1989), 5321–5399.

[17] Fuchs, J., Runkel, I., and Schweigert, C., TFT construction of RCFT correlators I: Partition functions. *Nucl. Phys. B* **646** (2002), 353–497.

[18] —, TFT construction of RCFT correlators II: Unoriented world sheets. *Nucl. Phys. B* **678** (2004), 511–637.

[19] —, TFT construction of RCFT correlators III: Simple currents. *Nucl. Phys. B* **694** (2004), 277–353.

[20] —, TFT construction of RCFT correlators IV: Structure constants and correlation functions. *Nucl. Phys. B* **715** (2005), 539–638.

[21] —, Ribbon categories and (unoriented) CFT: Frobenius algebras, automorphisms, reversions. Preprint; math.CT/0511590.

[22] Fukuma, M., Hosono, S., and Kawai, H., Lattice topological field theory in two dimensions. *Comm. Math. Phys.* **161** (1994), 157–176.

[23] Gawędzki, K., and Reis, N., WZW branes and gerbes. *Rev. Math. Phys.* **14** (2002), 1281–1334.

[24] Graham, K., and Watts, G. M. T., Defect lines and boundary flows. *J. High Energy Phys.* **0404** (2004), 019.

[25] Huang, Y.-Z., Vertex operator algebras, the Verlinde conjecture and modular tensor categories. *Proc. Natl. Acad. Sci. USA* **102** (2005), 5352–5356.

[26] Huang, Y.-Z., and Kong, L., Open-string vertex algebras, tensor categories and operads. *Comm. Math. Phys.* **250** (2004), 433–471.

[27] Kawahigashi, Y., Longo, R., and Müger, M., Multi-interval subfactors and modularity of representations in conformal field theory. *Comm. Math. Phys.* **219** (2001), 631–669.

[28] Lauda, A. D., Frobenius algebras and ambidextrous adjunctions. Preprint; math.QA/0502550.

[29] Longo, R., and Rehren, K.-H., Nets of subfactors. *Rev. Math. Phys.* **7** (1995), 567–598.

[30] Moore, G., K-Theory from a physical perspective. In *Topology, Geometry and Quantum Field Theory* (ed. by U. Tillmann), London Math. Soc. Lecture Note Ser. 308, Cambridge University Press, Cambridge 2004, 194–234.

[31] Moore, G., and Seiberg, N., Classical and quantum conformal field theory. *Comm. Math. Phys.* **123** (1989), 177–254.

[32] Müger, M., From subfactors to categories and topology I. Frobenius algebras in and Morita equivalence of tensor categories. *J. Pure Appl. Algebra* **180** (2003), 81–157.

[33] Nikshych, D., Turaev, V., and Vainerman, L., Quantum groupoids and invariants of knots and 3-manifolds. *Topology Appl.* **127** (2003), 91–123.

[34] Nikshych, D., and Vainerman, L., Finite quantum groupoids and their applications. In *New Directions in Hopf Algebras* (ed. by S. Montgomery and H.-J. Schneider), Math. Sci. Res. Inst. Publ. 43, Springer-Verlag, New York 2002, 211–262.

[35] Ostrik, V., Module categories, weak Hopf algebras and modular invariants. *Transform. Groups* **8** (2003), 177–206.

[36] Petkova, V. B., and Zuber, J.-B., Generalised twisted partition functions. *Phys. Lett. B* **504** (2001), 157–164.

[37] —, The many faces of Ocneanu cells. *Nucl. Phys. B* **603** (2001), 449–496.

[38] Reshetikhin, N. Yu., and Turaev, V. G., Ribbon graphs and their invariants derived from quantum groups. *Comm. Math. Phys.* **127** (1990), 1–26.

[39] Schellekens, A. N., and Yankielowicz, S., Simple currents, modular invariants, and fixed points. *Int. J. Mod. Phys. A* **5** (1990), 2903–2952.

[40] Segal, G. B., Topological structures in string theory. *Phil. Trans. Roy. Soc. London* **359** (2001), 1389–1398.

[41] Stevenson, M., The geometry of bundle gerbes, Ph.D. thesis, Adelaide 2000; math. DG/0004117

[42] Stolz, S., and Teichner, P., What is an elliptic object? In *Topology, Geometry and Quantum Field Theory* (ed. by U. Tillmann), London Math. Soc. Lecture Note Ser. 308, Cambridge University Press, Cambridge 2004, 247–343.

[43] Turaev, V. G., *Quantum Invariants of Knots and* 3*-Manifolds.* De Gruyter Stud. Math. 18, Walter de Gruyter, Berlin, New York 1994.

[44] Witten, E., Quantum field theory and the Jones polynomial. *Comm. Math. Phys.* **121** (1989), 351–399.

[45] Witten, E., On holomorphic factorization of WZW and coset models. *Comm. Math. Phys.* **144** (1992), 189–212.

[46] Yamagami, S., Frobenius algebras in tensor categories and bimodule extensions. *Fields Inst. Commun.* **43** (2004), 551–570.

Organisationseinheit Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany
E-mail: schweigert@math.uni-hamburg.de

Avdelning fysik, Karlstads Universitet, Universitetsgatan 5, 65188 Karlstad, Sweden
E-mail: jfuchs@fuchs.tekn.kau.se

Department of Mathematics, King's College London, Strand, London WC2R 2LS, England
E-mail: ingo@mth.kcl.ac.uk

# Soliton dynamics and scattering

Avy Soffer*

**Abstract.** A survey of results and problems of soliton dynamics in dispersive and hyperbolic nonlinear PDE's and the related spectral and scattering theory. I focus on the problem of large time behavior of the nonlinear Schrödinger equation, with both solitary and radiative waves appearing in the solution. The equations are nonintegrable in general and in arbitrary dimension. I will formulate the main conjectures relevant to soliton dynamics.

**Mathematics Subject Classification (2000).** Primary 35Qxx.

**Keywords.** Solitons, NLS, asymptotic-completeness, dispersive waves, asymptotic stability.

## 1. Introduction

The advances in spectral and scattering theory of the last 20 years, combined with the intense physical research based on nonlinear dispersive equations have led to major progress in nonlinear dynamics.

The goal of understanding the large time behavior of all solutions of nonlinear dispersive equations is now pursued on many levels: global existence theory, nonlinear scattering, new solutions, applications.

Furthermore, the new applications of NLPDE in physics generate a host of new research in the physics literature, e.g. experimental [12], [9], [30] and theoretical [57], [2], [29].

It is then fair to say that we are witnessing a new (golden) generation that focusses on complex collective systems behavior, the age of coherent structures.

Consider the following generic form of dispersive NLPDE:

$$i\,\partial_t u = Hu + F(u)u$$

with initial data in (say) some Hilbert space $\mathcal{H}$, typically (vector valued) Sobolev space.

$H$ is a self-adjoint (matrix) operator in general.

For example NLS (semilinear):

$$H = -\Delta, \quad F(u) \sim \sum_{i=1}^{N} \lambda_i |u|^{p_i},$$

$$\mathcal{H} = H^1(\mathbb{R}^n);$$

*Research partially supported by NSF Grant# DMS-0501043.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

and the NLKG equation (hyperbolic type):

$$\partial_t \begin{pmatrix} u \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ H_m & 0 \end{pmatrix} \begin{pmatrix} u \\ \dot{u} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ F(u) & 0 \end{pmatrix} \begin{pmatrix} u \\ \dot{u} \end{pmatrix},$$

$$H_m = -\Delta + m^2.$$

In general we expect that if the $\lambda_i > 0$ and the $p_i$ are not too large/small that global existence holds, that uniqueness holds and scattering theory holds and all solutions disperse like free waves for large time [54]. When some $\lambda_i$ are negative $F$ is attractive and blow-up occurs in general, in finite time, for data not too small.

However a fundamental new phenomena appears: the dispersion by $H$ can be exactly cancelled by the focusing of $F$, to create a new type of localized in space solutions, smooth, uniformly in time. I shall refer to these solutions as *coherent structures or solitons*.

**Example 1.1.** NLS solitons, KG/SG kinks, vortices, monopoles, breathers, topological solitons, hedgehogs, skyrmions, blackholes, ... and the more recent more exotic coherent structures: compactons, peakons, noncommutative solitons. [31], [54], [53].

**Definition 1.1.** A localized solution/soliton/coherent structure is a solution of a dispersive equation satisfying

$$\lim_{R\to\infty} \left\{ \sup_t \|\chi(|x| > R)u(x,t)\|_{\mathcal{H}} \right\} = 0.$$

To understand how fundamental are coherent structures we have the following conjecture:

**Grand Conjecture** (Asymptotic Completeness). Generic asymptotic states are given by independently (freely) moving coherent structures and free radiation.

**Comments.** 1) The Grand Conjecture states that besides free waves only these coherent structures can emerge as $t \to \infty$.

2) We are very far from proving such a result for any interesting equation.

In making progress in this direction I will formulate another "simple" conjecture that I expect to play an important role:

**Petite Conjecture.** Localized solutions of NLS are almost periodic in time.

**Comments.** 1) I used "NLS" and not "dispersive" to avoid giving a rigorous definition of "dispersive".

2) This conjecture is a nonlinear analog of the geometric characterisation of bound states in linear theory originally proposed by Ruelle and developed as RAGE theorem.

It also states that coherent structures/solitons are the "bound states" of dispersive wave equations.

3) While such a result follows for data near a stable soliton solution of NLS [40], [37], [35], [10], [45], [46], [4], [5], [36] it is not known for even a single equation in such generality.

4) One can replace the localization assumption by the stronger condition

$$\sup_t \|\chi(|x| > R)u(x,t)\|_{\mathcal{H}} \leq cR^{-m}, \quad R > 1,$$

some $m > 0$, or even exponential decay in $R$ for many equations.

## 2. Asymptotic stability

The kind of problems we do understand are small perturbations of the putative asymptotic states described above. That is, we can prove that if the initial data is close to $N$ solitons moving independently and small perturbation it will propagate as expected when $t \to \infty$.

Next, I shall describe the developments and arguments leading to this result.

So consider the NLS in three or more dimensions, here we follow [40]:

$$i\frac{\partial \psi}{\partial t} = -\Delta\psi - F(|\psi|^2)\psi \quad x \in \mathbb{R}^n \text{ and } n \geq 3. \tag{1}$$

If $F$ has a negative (attractive) part the equation will have, in general, coherent structure solutions, such as solitons. To find (some of) them, we look for time periodic solutions

$$\psi = e^{i\omega t}\phi_\omega(x),$$

which gives

$$-\omega\phi_\omega = -\Delta\phi_\omega - F(|\phi_\omega|^2)\phi_\omega.$$

In general $\phi_\omega$ will be localized (at least as an $L^2$-function) for $\omega > 0$.

Thanks to the pioneering works of [8], [53], [3], [24], [32], [17] we know a great deal of information about such solitons: existence, decay at infinity, uniqueness of the positive solutions, symmetry and more.

By Galilean invariance

$$\psi_\sigma \equiv e^{i\vec{v}\cdot x - i\frac{1}{2}(|v|^2-\omega)t + i\gamma}\phi_\omega(x - \vec{v}t - \vec{a})$$

are all solutions of NLS;

$$\sigma = (\vec{u}, \gamma, \vec{a}, \omega) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}.$$

$\psi_\sigma$ are moving solitons, with velocity $\vec{v}$ localized at $\vec{a}$ (at time 0), with energy $\omega$ and phase $\gamma$. The general conjecture then states that generic solutions of NLS are

asymptotic to a sum of such moving solitons plus a free wave (a solution of the free Schrödinger equation corresponding to $F = 0$).

A weaker but useful result is orbital *stability* of solitons: we would like to know if a small perturbation of a soliton (as initial data) leads to a nearby, soliton (in $\sigma$ space) up to a phase, for all times [7], [13], [14], [18], [19], [20], [58], [59].

It turns out that in many cases stability follows from *linear stability* :

First we linearize the equation around a soliton:

$$\psi \equiv e^{i\theta}(\phi_\omega + R)$$

and deriving (the linear part of) the equation of $R$: since there will be both $R$ and $\bar{R}$ terms in the equation we complexify to get:

$$i\frac{\partial}{\partial t}\begin{pmatrix} R \\ \bar{R} \end{pmatrix} = \mathcal{H}\begin{pmatrix} R \\ \bar{R} \end{pmatrix},$$

$$\mathcal{H} \equiv \begin{pmatrix} L+ & W(x) \\ -W(x) & -L_+ \end{pmatrix},$$

$$L_+ = -\Delta + \omega - F(\phi_\omega^2) - F'(\phi_\omega^2)\phi_\omega^2,$$

$$W(x) = -F'(\phi_\omega^2)\phi_\omega^2.$$

Since $\mathcal{H}$ is not self-adjoint we do not have $L^2$ conservation under such $\mathcal{H}$. Linear stability states that on the complement of the root space of $\mathcal{H}$ the solutions are uniformly bounded in $L^2$:

$$\sup_t \|U(t)Pf\|_{L^2} \leq c\|f\|_{L^2}$$

where $P$ is the projection on $(N^*)^\perp$ and $N$ is the root space of $\mathcal{H}$,

$$N \equiv \bigcup_{\ell \geq 1} \ker \mathcal{H}^\ell.$$

Such analysis can be made for other coherent structures in a similar way [vortices, kinks, blackholes,. . . .].

To proceed with the problem of asymptotic stability and completeness we are going to need much more detailed spectral/scattering results for $e^{i\mathcal{H}t}P$. Proving such results for various types of $\mathcal{H}$ that appear in applications is in general difficult, and while there are many works in this direction from the last few years, a lot is left open, see the review [42].

We need to prove

a) Absence of embedded eigenvalues in the continuous spectrum of $\mathcal{H}$.

There are remarkably detailed and complete results for self-adjoint Schrödinger operators on $\mathbb{R}^n$ and manifolds [22], [41].

The extension to matrix operators such as $\mathcal{H}$ is still eluding us, and a general new approach would be of great interest to the field, see however [42].

b) Absence of threshold resonances. This is related to the next condition.

c) Dispersive estimates:

$$\|U(t)P\psi_0\|_{L^\infty} \lesssim t^{-n/2}\|\psi_0\|_{L^1}.$$

Weaker estimates are also of interest, such as

$$\|\langle x\rangle^{-\sigma}U(t)P\langle x\rangle^{-\sigma}\| \lesssim t^{-\sigma}, \quad t > 1.$$

The subject of dispersive estimates is another fast growing field of research inspired by soliton dynamics, see the latest review of Schlag [42].

Such $L^p$-dispersive estimates were proved for

$$H = -\Delta + V \text{ on } \mathbb{R}^n, \quad n \geq 3,$$

for a quite general class of $V$ in [23]. It was then extended (also to $L^p$ boundedness of wave operators) in [60]; further important new extension is due to [38]. For matrix operators see [39], [42]. All the results which are known require enough decay at $\infty$ of the potential perturbations. The decay is determined by the decay of the soliton.

NLS solitons decay exponentially and so they are included.

However other types of coherent structures like vortices, monopoles etc. lead to slowly decaying potentials, and so the dispersive theory for both Schrödinger and matrix operators is still lacking the generality needed.

What is more, the above stated decay also requires the knowledge of absence of threshold resonances, a difficult problem to solve for specific operators. In fact, in some situations (e.g. kink scattering) such resonances are present, and therefore the asymptotic stability theory of kinks requires as yet unknown modifications.

## 3. The method of modulation equations

Presently, the modulation equations approach is the only method that is used to prove asymptotic stability of solitons. The first step is the *Ansatz* and the associated *orthogonality condition*.

**Ansatz.**

$$\psi(t) = S_{\sigma(t)} + R(t).$$

Here $S_\sigma$ is a soliton with parameters $\sigma$.

**Orthogonality condition.** $R(t)$ belongs to the range of $P_t$ so it is orthogonal to the root space adjoint $N^*$.

**Remark 3.1.** When the soliton is small we can replace $P_t$ by $P_{t=0}$ [45], [46]. Under favorable conditions we can derive a closed system of equations for $R(t), \sigma(t)$. The equation for $\mathbb{R} \equiv (R(t), \bar{R}(t))$ then has the form

$$i\partial_t \mathbb{R}(t) = \mathcal{H}\mathbb{R}(t) + \mathcal{NL}(\mathbb{R}(t))$$

where $\mathcal{NL}(\mathbb{R}(t))$ is a sum of quadratic and higher order terms in $\mathbb{R}$, and terms of the form

$$O(\dot{\sigma}\mathbb{R}).$$

The $\sigma$-equation is a system of ODE's:

$$\dot{\sigma}(t) = \mathcal{K}(t)\sigma(t) + \mathcal{NL}(\mathbb{R}, \sigma).$$

The idea for solving such a system is based on bootstrap and smallness. Suppose that we can prove that the solutions of the PDE for $(\mathbb{R}(t))$ is indeed dispersive:

$$|||\mathbb{R}(t)||| = O(t^{-m})$$

where $|||\cdot|||$ is typically a sum of $L^P$ and local decay norms.

For example, for $p = \infty$, and three or more dimensions we expect $m = 3/2$ at least. Then, plugging this estimate for $\mathbb{R}(t)$ into the $\mathcal{NL}(\mathbb{R}, \sigma)$ terms of the ODE we shall be able to prove that

$$\dot{\sigma}(t) \text{ is in } L^1(dt) \text{ (excluding } \dot{\gamma})$$

and perhaps get some pointwise decay.

Then these estimates on $\dot{\sigma}$ should be sufficient to control the $O(\dot{\sigma}R)$ terms in the PDE. Smallness can then be used to close the above self-consistent system of estimates.

The key to implementing this procedure is proving sufficient decay estimates for the PDE, assuming the decay of $\dot{\sigma}$. This leads us to study the following general problem:

$$i\frac{\partial R}{\partial t} = \mathcal{H}PR + P\mathcal{NL}(R)$$

where $P$ is the projection on $(N^*)^{\perp}$.

When $R$ is small, we try to solve this problem by Duhamel type identities

$$R(t) = e^{-i\mathcal{H}Pt}R(0) - i\int_0^t e^{-i\mathcal{H}P(t-s)}P\mathcal{NL}(R(s)\,ds.$$

The starting point is then knowing the basic estimates for

$$e^{-i\mathcal{H}Pt}$$

which is the linear problem described before.

There are many situations where the range of $P$ includes, besides the continuous spectrum of $\mathcal{H}$, some point spectrum. In this case, further decomposition of the solution is needed.

This will be described in the next section.

So consider the case when $P = P_c(\mathcal{H})$. We need, besides the $L^p$ decay estimates for $\mathcal{H}$, to estimate the nonlinear terms.

When the nonlinearities are of sufficient power decay (near zero) this is not difficult.

However, notice that we always need to deal with a term of the form

$$\sum SR^2.$$

When the soliton $S$ is well localized, local decay pointwise in time plus some $L^p$ decay is sufficient. The complications arise in cases where there is not enough decay or when $S$ is not well localized. For example the kink problem leads to $S$ that is of order 1 at infinity! Hence all the decay should come from $R^2$, which is not possible in 1-dimension. For this and other equally subtle reasons the kink asymptotic stability is completely open. In particular, one needs to solve the *long range* nonlinear scattering problem. A lot of work was done on this problem by Ginibre–Velo–Naumkin [16], [21] and collaborators. Recently Delort [11] worked out the problem for NLKG in 1-dimension. A new simpler and general approach was very recently developed in [26], [27], [28] ]

In the simplest cases we end up with estimating

$$R(t) \sim c^{-i\mathcal{H}_c t} R(0) - i \int_0^t e^{-i\mathcal{H}_c(t-s)} O(\chi R^2(s) + R(s)^p + O(s^{-\alpha})R(s))ds$$

in $L^p$.

A particularly effective way of doing it, is to use the mixed norm $L^\infty + L^2$.

We then get

$$\|R(t)\|_{L^2+L^\infty} \leq C\|R(0)\|_{L^2+L^1}\langle t\rangle^{-n/2} + c\int_0^t \frac{ds}{\langle t-s\rangle^{n/2}}\|O(\cdot)\|_{L^2+L^1}\,ds.$$

In three or more dimensions, $n/2 > 1$.

Since for all $s$ large enough and any $\varepsilon > 0$,

$$\begin{aligned}&\|\mathcal{O}(\cdot)\|_{L^2+L^1}\\ &\quad\leq C\langle s\rangle^{-n/2}\sup_{0\leq s'\leq s}\{\|R(s')\|_{2+\infty}^2 + \|R(s')\|_{2+\infty}^{2+m} + \|R(s')\|_p^p + \varepsilon\|R(s')\|_{2+\infty}\}\end{aligned}$$

provided $\left(\frac{n}{2} - \frac{n}{p}\right)\cdot p \geq \frac{n}{2}$ implies $p \geq 3$. $\|R(t)\|_{L^2+L^\infty}$ is $O(t^{-n/2})$ for $R(o)$ small.

It is possible to improve the estimates to lower values of $p$, by proving only that $\|R\|_q \leq ct^{-1-\varepsilon}$ for

$$q \text{ such that } \frac{n}{2} - \frac{n}{q} = 1 + \varepsilon.$$

A different approach, which applied to nonlocalized perturbations of solitons is based on using Strichartz estimates instead of pointwise bounds. See [15].

## 4. Selection of the ground state, spurious eigenvalues

In general there is more than one (family of) localized states to the nonlinear equation.

Examples include

$$i\frac{\partial\psi}{\partial t} = (-\Delta + V(x))\psi + \lambda F(|\psi|)\psi \tag{2}$$

with $H = -\Delta + V(x)$ having more than one (negative) eigenvalue in its spectrum. Another example is the nonlinear wave equation analog of (2):

$$(\partial_t^2 - \Delta + V(x) + m^2)u = \lambda F(u)u \tag{3}$$

with $-\Delta + V(x) + m^2 > 0$, and having as before one or more eigenvalues.

Other examples include the NLS with excited state solitons etc., spurious eigenvalues, including embedded eigenvalues in the spectrum of the linearization. In these cases the grand conjecture is more complicated to state.

The generic behavior we expect is that the asymptotic states are combinations of ground state (families) solitons *only* plus free radiation.

This was proved for the NLWE (3) in the case of one bound state in 3-dimensions in [49]. The techniques developed in [49] will be briefly described below; they were used to deal with the more involved NLS (2) in the case of two bound states in [51], [56] and some results were obtained for more then two bound states in [55]. It also applied to the *linear* resonance problem in QM [47], [48], [50], [33], [34], [18], [6]. Very recently an experiment was done confirming the predictions below [30], [52].

We use modulation equations again. Consider the problem (2): Time periodic soliton solutions, nonlinear bound states, bifurcate from the linear eigenstates of $-\Delta + V(x)$. For each $E_i$, $i = 0, 1$, we solve

$$(-\Delta + V(x))\psi_{E_i} + \lambda F(|\psi_{E_i}|^2)\psi_{E_i} = E_i\psi_{E_i} \tag{4}$$

such that $E_i \to E_{i*}$ as $\lambda \to 0$, where $(i = 0, 1)$

$$(-\Delta + V(x))\psi_i = E_{*_i}\psi_i \ \|\psi_i\| = 1.$$

We assume the initial data is small in $H^s$, $(s \geq 2)$ and $2E_{1*} - E_{0*} > 0$.

We begin with the Ansatz $\phi(t) \equiv e^{-i\theta(t)}[\psi_0(t) + \psi_1(t) + \phi_2(t)]$ where $\psi_0(t) \equiv \psi_{E_0(t)}$ is a solution of the ground state eigenvalue equation with energy $E_0(t)$, at time $t$. $E_0(t)$ will be determined later by orthogonality conditions [Se 7, 5, 1]. Similarly $\psi_1(t)$ is an excited state eigenvector with eigenvalue $E_1(t)$. $\theta(t) \equiv \theta_0(t) + \tilde{\theta}(t)$; $\theta_0(t) = \int_0^t E_0(s)ds$. $\tilde{\theta}(t)$ will be chosen appropriately; it includes (logarithmic) divergent phase. Substitution of the above Ansatz for $\phi$ into (1), and complexifying the equations $[\phi_2 \to (\phi_2, \bar{\phi}_2) \equiv \Phi_2(t)$, $\psi_j \to (\psi_j, \bar{\psi}_j) \equiv \Psi_j(t)$ etc.] we derive

$$i\partial_t\Phi_2(t) = \mathcal{H}_0(t)\Phi_2(t) - i\partial_t\Psi_0 - \left[((E_0 - E_1) + \partial_t\tilde{\theta})\sigma_3 + i\partial_t\right]\Psi_1 + \vec{F}_{NL},$$

where $\vec{F}_{\mathcal{NL}}$ is nonlinear in $\Phi_2$, $\Psi_0$, $\Psi_1$, $\tilde{\theta}$ and $\mathcal{H}_0(t)$ is given by the matrix operator

$$\sigma_3 \begin{pmatrix} H - E_0(t) + 2\lambda|\psi_0(t)|^2 & \lambda\psi_0^2(t) \\ \lambda\bar{\psi}_0^2(t) & H - E_0(t) + 2\lambda|\psi_0(t)|^2 \end{pmatrix} \tag{5}$$

where $\sigma_3$ is the Pauli matrix diag$(1, -1)$. We consider the spectrum of $\mathcal{H}_0(t)$ for fixed $t$, and $|\psi_0| \equiv |\alpha_0|$ small: (a) The continuous spectrum extends from $-\mu$ to $-\infty$, and $\mu$ to $\infty$ where $\mu \equiv E_1 - E_0 + O(|\alpha_0|^2)$. The discrete spectrum is $\{0, -\mu, \mu\}$, with $0 < |\mu| < |E_0|$ by assumption. (b) Zero is a generalized eigenvalue of $\mathcal{H}_0$, with generalized eigenspace spanned by $\{\sigma_3\Psi_0, \partial_{E_0}\Psi_0\}$.

The discrete spectral subspace has dimension four. Therefore, $\Psi_2$ which lies in the continuous spectral part of $\mathcal{H}_0(t)$, is constrained by four orthogonality conditions. Furthermore $\partial_t\tilde{\theta}$ is chosen to remove divergent logarithmic phase contributions. In the weakly nonlinear (perturbative) regime, bound states have expressions $\psi_{E_j} = \alpha_j(\psi_{j*}(x) + g|\alpha_j|^2\psi_j^{(1)}(x) + \mathcal{O}(g^2|\alpha_j|^4))$ and $E_g = E_{j*} + \mathcal{O}(|\alpha_j|^2)$. The system for $\Phi_2$ and $\vec{\alpha} = (\vec{\alpha}_0, \vec{\alpha}_1)$ can be written in the form $i\partial_t\vec{\alpha} = \mathcal{A}(t)\vec{\alpha} + F_\alpha$, $i\partial_t\Phi_2 = \mathcal{H}(t)\Phi_2 + F_\Phi$ .

To proceed further we decompose $\Phi_2$ into its continuous spectral (dispersive) part, $\eta \in \mathcal{H}_0(T)$, and its components along the discrete modes. The latter are higher order and controllable. Thus NLS at low energy is equivalent to a system of the form:

$$\begin{aligned} i\partial_t\eta &= \mathcal{H}_0(T)\eta + \mathcal{F}_\eta(t; \alpha_0, \beta_1, \eta) + \sigma(t)\eta, \\ i\partial_t\beta_1 &= 2\lambda\langle\psi_{0*}, \psi_{1*}^3\rangle|\beta_1|^2\alpha_0 e^{i\lambda_+ t} \\ &\quad + 2\lambda\langle\psi_{0*}\psi_{1*}^2, \pi_1\Phi_2\rangle\bar{\beta}_1\alpha_0 e^{2i\lambda_+ t} + \mathcal{R}_0, \\ i\partial_t\alpha_0 &= \lambda\langle\psi_{0*}^2, \psi_{1*}^2\rangle e^{-2i\lambda_+ t}\beta_1^2\overline{\alpha}_0 \\ &\quad + \lambda\langle\psi_{0*}\psi_{1*}^2, \Phi_2\rangle\beta_1^2 e^{-2i\lambda_+ t} + \mathcal{R}_1, \end{aligned} \tag{6}$$

where $\mathcal{R}_j$ denotes corrections of a similar form and higher order.

The above system can be viewed as an infinite dimensional Hamiltonian system consisting of two subsystems: a finite dimensional subsystem governing "oscillators", $(\alpha_0, \beta_1)$, and an infinite dimensional subsystem governing the field, $\eta$.

The coupled system (6) can not be solved or understood by looking at the linear terms only. This is due to the fact that the asymptotic behavior is determined by a process, nonlinear, in which the excited state part of the solution decays into radiation and ground state part. We therefore need to derive effective equations for the ground and excited state parts, which include the dissipative effects due to coupling to radiation. see [49], [44] see also [43], [1], [25].

To arrive at the reduction, we solve the $\eta$-equation, making explicit all terms through second order in $g$, using the Green's function $G(t, t') = e^{-i\mathcal{H}_0(T)(t-t')}$. We focus on the key terms coming from the sources in $\mathcal{F}_\eta$ or the type $\alpha_0^i\alpha_1^j$, $0 \le i, j \le 2$

and having oscillatory phases $e^{im_{ij}t}$. Their contribution to $\eta$ is of the form

$$\sim \int_0^t e^{-i\mathcal{H}_0(T)(t-t')}|\chi\rangle e^{im_{ij}(t')}\alpha_0^i(t')\alpha_1^j(t')dt'$$

where $\alpha_0$, $\alpha_1$ is a component of either $\vec{\alpha}_0$ or $\vec{\alpha}_1$, where $|\chi\rangle$ is an (exponentially localized) function of position expressible in terms of $\psi_{0*}$ and $\psi_{1*}$. We insert this solution into the $\alpha_0$-, $\alpha_1$-equations, in place of $\Phi_2$. We obtain integro-differential equations for $\alpha_0$, $\alpha_1$, $(\beta_1)$. The resulting terms of the above form are solutions to a forced linear system and among the forcing terms there are (coupled) oscillatory terms with the frequency $\omega_*$, which is resonant with the continuous spectrum. Internal dissipation resulting in nonlinear resonant energy transfer from the excited state to the ground state and to dispersive radiation is derived from these resonant terms; see also the derivation of internal dissipation in both linear and nonlinear resonance theories recently developed by us [46], [44]. This dissipation coefficient is $\Gamma$, the rate of decoherence and relaxation. The above described scheme gives $i\partial_t\tilde{\alpha}_0 = (-\Lambda + i\Gamma) \times |\tilde{\beta}_1|^2\tilde{\alpha}_0 + \tilde{\mathcal{R}}_0(t)$, $i\partial_t\tilde{\beta}_1 = 2(\Lambda - i\Gamma)|\tilde{\alpha}_0|^2|\tilde{\beta}_1|^2\beta_1 + \tilde{\mathcal{R}}_1(t)$.

Introducing the squared projections of the system's state onto the ground state and excited states, $P_0 \equiv |\tilde{\alpha}_0|^2$, $P_1 \equiv |\tilde{\beta}_1|^2$ we obtain NLME. The system is analyzed in terms of renormalized powers $\mathcal{Q}_0$ and $\mathcal{Q}_1$, for which it is shown that there exist transition times $t_0$ and $t_1$, such that $\mathcal{Q}_0(t)$ decays rapidly on $[0, t_0]$, $\mathcal{Q}_0(t)/\mathcal{Q}_1(t)$, grows rapidly on $[t_0, t_1]$, and then finally on $[t_1, \infty)$ the following system governs: $\partial_t\mathcal{Q}_0 = 2\Gamma\mathcal{Q}_0\mathcal{Q}_1^2$, $\partial_t\mathcal{Q}_1 = -4\Gamma\mathcal{Q}_0Q_1^2$. This gives $Q_0 \uparrow Q_0(\infty)$ and $Q_1 \downarrow 0$ at rates discussed above.

$$\Gamma = \pi\lambda^2|(e_{\omega*}, \psi_{1*}^2\psi_{0*})|^2$$

for $F(|\psi|^2) \equiv |\psi|^2$. Here $e_{\omega*}$ is the generalized eigenvalue of $H_0$ at energy $\omega* = 2E_{1*} - E_{0*}$.

## 5. Concluding remarks

Nonlinear dispersive equations play a prominent role in many fields of physics, including BEC theory, nonlinear optics, large molecule dynamics (e.g. DNA) and more.

The mathematical aspects of such equations is remarkably rich, and contributed new, challenging directions for PDE, mathematical physics spectral and scattering theory and more. In particular, the problem of large time dynamics of interacting solitons and more general coherent structures is witnessing a major progress in the last 15–20 years. Yet, we are only at the beginning of developing the mathematical tools and theories to deal with the general aspects of soliton dynamics. Besides the many problems I listed in the previous sections, other topics worth mentioning are: systems of equations (e.g. coupled Maxwell–Dirac equations and BEC coupled to vapor at finite temperature), solitons dynamics on curved spaces, like arbirarily shaped optical fibre (monoples and other topological solitons), and discrete space modeles.

## References


[1] Allen, L., Eberly, J. H., *Optical resonance and two-level atoms*. Dover, 1987.

[2] Aschbacher, W. H. et al., Symmetry breaking regime in the nonlinear Hartree equation. *J. Math. Phys.* (*N.Y.*) **43** (2002), 3879–3891.

[3] Berestycki, H., Lions, P. L., Existence d'ondes solitaires dans les problemes nonlineares du type Klein-Gordon. *C. R. Acad. Sci. Paris* **288** (7) (1979), 395–398.

[4] Buslaev, V. S., Perelman, G. S., Scattering for the nonlinear Schrödinger equation: states that are close to a soliton. *Algebra i Analiz* **4** (6) (1992), 63–102; English transl. *St. Petersburg Math. J.* **4** (6) (1993), 1111–1142.

[5] Buslaev, V. S., Perelman, G. S., On the stability of solitary waves for nonlinear Schrödinger equations. In *Nonlinear evolution equations*, Amer. Math. Soc. Transl. (2) 164, Amer. Math. Soc., Providence, RI, 1995, 75–98.

[6] Cattaneo, L., Graf, G. H., Hunziker, W., A general resonance theory based on Mourre's inequality. arXive math-ph/0507063.

[7] Cazenave, T., Lions, P.-L., Orbital stability of standing waves for some nonlinear Schrödinger equations. *Comm. Math. Phys.* **85** (1982), 549–561.

[8] Coffman, C. V., Uniqueness of positive solutions of Laplace $u-u+u^3 = 0$ and a variational characterization of other solutions. *Arch. Rat. Mech. Anal.* **46** (1972), 81–95.

[9] Davis, K. B., et al., Bose-Einstein Condensation in a Gas of Sodium Atoms. *Phys. Rev. Lett.* **75** (1995), 3969.

[10] Deift, P., Zhou X., *Long time behavior of the non-focusing Schrödinger equation – a case study*. Lectures in Mathematical Sciences 5, University of Tokyo, 1994.

[11] Delort, J.-M., Global solutions for small nonlinear long range perturbations of two dimensional Schrödinger equations. *Mém. Soc. Math. France* **91** (2002), 94 pp.

[12] Ensher, J. R., et al., Bose-Einstein Condensation in a Dilute Gas: Measurement of Energy and Ground-State Occupation. *Phys. Rev. Lett.* **77** (1996), 4984–4987.

[13] Fröhlich, J., Gustafson, S., Jonson, B. L. G., Sigal, I. M., Solitary wave dynamics in an external potential. *Comm. Math. Phys.* **250** (2004), 613–642.

[14] Fröhlich, J., Tsai, T. P., Yau, H. T., Dynamics of solitons in the nonlinear Hartree equation. Preprint.

[15] Gustafson, K., Nakanishi, K., and Tsai, T. P., Asymptotic stability and completeness in the energy space for nonlinear Schrödinger equations with small solitary waves. *Internat. Math. Res. Notices* **2004** (66) (2004), 2559–3589.

[16] Ginibre, J., Velo, G., Long-Range scattering and modified wave operators for some Hartree type equations: III. Gevrey spaces and low dimensions. *J. Differential Equations* **175** (2001), 415–501.

[17] Gidas B., Ni, W. M., Nirenberg L., Symmetry and related properties via the maximum principle. *Comm. Math. Phys.* **68** (3) (1979), 209–243.

[18] Grillakis, M., Analysis of the linearization around a critical point of an infinite dimensional Hamiltonian system. *Comm. Pure Appl. Math.* **41** (6) (1988), 747–774.

[19] Grillakis, M., Shatah, J., Strauss, W., Stability theory of solitary waves in the presence of symmetry. I. *J. Funct. Anal.* **74** (1987), 160–197.

[20] Grillakis, M., Shatah, J., Strauss, W., Stability theory of solitary waves in the presence of symmetry. II. *J. Funct. Anal.* **94** (1990), 308–348.

[21] Hayashi, N., Naumkin, P., Asymptotics for large time of solutions to the nonlinear Schrödinger and Hartree equations. *Amer. J. Math.* **20** (1998), 369–389.

[22] Ionescu, A., Jerison, D., On the absence of positive eigenvalues of Schrödinger operators with rough potentials. *Geom. Funct. Anal.* **13** (2003), 1029–1081.

[23] Journé, J.-L., Soffer, A., Sogge, C. D., Decay estimates for Schrödinger operators. *Comm. Pure Appl. Math.* **44** (5) (1991), 573–604.

[24] Kwong, M. K., Uniqueness of positive solutions of Laplace $u - u + u^p = 0$ in $R^n$. *Arch. Rat. Mech. Anal.* **65** (1989), 243–266.

[25] Lamb, H., On a peculiarity of the wave-system due to the free vibrations of a nucleus in an extended medium. *Proc. London Math. Soc.* **32** (1900), 208–211.

[26] Lindblad, H., Soffer, A., A remark on long range scattering for the nonlinear Klein-Gordon Equation. *J. Hyperbolic Differ. Equ.* **2** (1) (2005), 77–89.

[27] Lindblad, H., Soffer, A., A remark on asymptotic completeness for the critical nonlinear Klein-Gordon equation. *Lett. Math. Phys.* **73** (2005), 249–258.

[28] Lindblad, H., Soffer, A., Scattering and small data completeness for the critical nonlinear Schrödinger equation. *Nonlinearity* **19** (2006), 345–353.

[29] Lieb, E. H., et al., A rigorous derivation of the Gross-Pitaevskii energy functional for a two-dimensional Bose gas. *Comm. Math. Phys.* **224** (2001), 17.

[30] Mandelik, D., et al., Nonlinearly induced relaxation to the ground state in a two-level system. *Phys. Rev. Lett.* **95** (2005), 073902.

[31] Manton, N., Sutcliffe, P., *Topological Solitons.* Cambridge Monogr. Math. Phys., Cambridge University Press, Cambridge 2004.

[32] McLeod, K., Serrin, J., Uniqueness of positive radial solutions of $\Delta u + f(u) = 0$ in $\mathbb{R}^n$. *Arch. Rat. Mech. Anal.* **99** (1987), 115–145.

[33] Merkli, M., Sigal, I. M., A time dependent Theory of Quantum Resonances. *Comm. Math. Phys.* **201** (1999), 549–576.

[34] Miller, P., Soffer, A., and Weinstein, M., Methastability of breather modes. *Nonlinearity* **13** (2002), 507–568.

[35] Pego, R. L., Weinstein, M. I., Asymptotic stability of solitary waves. *Comm. Math. Phys.* **164** (1994), 305–349.

[36] Perelman, G., Asymptotic stability of multi-soliton solutions for nonlinear Schrödinger equations. *Comm. Partial Differential Equations* **29** (2004), 1051–1095.

[37] Pillet, C. A., Wayne, C. E., Invariant manifolds for a class of dispersive, Hamiltonian, partial differential equations. *J. Differential Equations* **141** (2) (1997), 310–326.

[38] Rodnianski, I., Schlag, W., Time decay for solutions of Schrödinger equations with rough and time-dependent potentials. *Invent. Math.* **155** (2004), 451–513.

[39] Rodnianski, I., Schlag, W., Soffer, A., Dispersive analysis of charge transfer models. *Comm. Pure Appl. Math.* **58** (2005), 149–216.

[40] Rodnianski, I., Schlag, W., Soffer, A., A symptotic stability of $N$-soliton states of NLS. Submitted.

[41] Rodnianski, I., Tao, T., Quantiative limiting absorption principles on manifolds, and applications. Preprint.

[42] Schlag, W., Dispersive Estimates for Schrödinger Operators: A survey. Preprint, 2005.

[43] Sigal, I. M., Nonlinear wave and Schrödinger Equations I. Instability of time periodic and quasi periodic solutions. *Comm. Math. Phys.* **153** (1993), 297.

[44] Soffer, A., Dissipation through dispersion. In *Nonlinear Dynamics and Renormalization Group* (ed. by I. M Segal and C. Sulen), CRM Proc. Lecture Notes 27, Amer. Math. Soc., Providence, RI, 2001, 175–184.

[45] Soffer, A., Weinstein, M., Multichannel nonlinear scattering for nonintegrable equations. *Comm. Math. Phys.* **133** (1990), 119–146.

[46] Soffer, A., Weinstein, M., Multichannel nonlinear scattering, II. The case of anisotropic potentials and data. *J. Differential Equations* **98** (1992), 376–390.

[47] Soffer, A., Weinstein, M., Nonautonomous Hamiltonians. *J. Statist. Phys.* **93** (1998), 359–391.

[48] Soffer, A., Weinstein, M., Time Dependent Resonance Theory. *Geom. Funct. Anal.* **8** (1998), 1086–1128.

[49] Soffer, A., Weinstein, M., Resonances, Radiation Damping and Instability in Hamiltonian nonlinear wave equation. *Invent. Math.* **136** (1999), 9–74.

[50] Soffer, A., Weinstein, M., Ionization and scattering for short lived potentials. *Lett. Math. Phys.* **48** (4) (1999), 339–352.

[51] Soffer, A., Weinstein, M., Selection of the Ground state for nonlinear Schrödinger Equations. *Rev. Math. Phys.* **16** (2004), 977–1071.

[52] Soffer, A., Weinstein, M., Theory of Nonlinear Dispersive Waves and Selection of the Ground state. *Phys. Rev. Lett.* **95** (2005), 213905.

[53] Strauss, W., Existence of solitary waves in higher dimensions. *Comm. Math. Phys.* **55** (1977), 149–162.

[54] Sulem, C., Sulem, P.-L., *The nonlinear Schrödinger equation. Self-focusing and wave collapse.* Appl. Math. Sci. 139, Springer-Verlag, New York 1999.

[55] Tsai, T., Asymptotic dynamics of nonlinear Schrödinger Equation with many bound states. *J. Differential Equations* **192** (2003), 225–282.

[56] Tsai, T. P., Yau, H. T., Stable directions for excited states of nonlinear Schrödinger equations. *Comm. Partial Differential Equations* **27** (2003), 2363–2402; Relaxation of excited states in nonlinear Schrödinger equations. *Internat. Math. Res. Notices* **2002** (31) (2002), 1629–1673.

[57] Ueda, M., Leggett, A. L., Macroscopic quantum tunneling of a Bose-Einstein condensate with attractive interaction *Phys. Rev. Lett.* **80** (1998), 1576.

[58] Weinstein, Michael I., Modulational stability of ground states of nonlinear Schrödinger equations. *SIAM J. Math. Anal.* **16** (3) (1985), 472–491.

[59] Weinstein, Michael I., Lyapunov stability of ground states of nonlinear dispersive evolution equations. *Comm. Pure Appl. Math.* **39** (1) (1986), 51–67.

[60] Yajima, K., The $W^{k,p}$-continuity of wave operators for Schrödinger operators. *J. Math. Soc. Japan* **47** (3) (1995), 551–581.

Mathematics Department, Rutgers, The State University, 110 Frelinghuysen Road,
Piscataway, NJ 08854-8019, U.S.A.
E-mail: soffer@math.rutgers.edu

# Hypocoercive diffusion operators

Cédric Villani

**Abstract.** In many problems coming from mathematical physics, the association of a degenerate diffusion operator with a conservative operator may lead to dissipation in all variables and convergence to equilibrium. One can draw an analogy with the well-studied phenomenon of hypoellipticity in regularity theory, and actually both phenomena have been studied together. Now a distinctive theory of "hypocoercivity" is starting to emerge, with already some striking results, and several challenging open problems.

**Mathematics Subject Classification (2000).** Primary 35B40; Secondary 35K70, 76P05.

**Keywords.** Hypocoercivity, hypoellipticity, diffusion equations, spectral gap, logarithmic Sobolev inequalities, Fokker–Planck and Boltzmann equations, $H$ Theorem.

## Introduction

During the past decade, considerable progress has been achieved in the qualitative study of diffusion equations in large time, be it for linear or nonlinear models. Quantitative functional methods have become especially popular. Here are some of the keywords in the field: spectral gap (Poincaré) inequalities, logarithmic Sobolev inequalities, analysis of entropy production, gradient flows, rescalings. Most of the time, estimates on the rate of convergence are established in the end by means of some Gronwall-type inequality $dE/dt \leq -\Phi(E)$, where $E$ is a Lyapunov functional for the system. Among a large literature, I shall only quote some of my own works: entropy production estimates for the spatially homogeneous Boltzmann equation, in collaboration with Giuseppe Toscani [25], [26], [28]; and for certain nonlinear diffusion equations with a convex mean-field interaction, in collaboration with José Antonio Carrillo and Robert McCann [2].

While these subjects are still very active, in this text I shall focus on a newer direction of research which has emerged only a few years ago, and can be loosely described as "the role of the non-dissipative part in the dissipation process".

Indeed, it happens not so rarely that the dissipative properties of an equation are strongly influenced by some of the conservative terms in this equation. This statement in itself is nothing new, since it is almost obvious in the context of hydrodynamics (dissipativity in Navier–Stokes is certainly considerably more complex than in the heat equation). In the context of diffusion equations, the interaction between dissipative and conservative terms is also well-known, since it is at the basis of the phenomenon of

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

*hypoellipticity*. To make the discussion a bit more precise, let me recall a particularly simple theorem of hypoelliptic regularization, which is a direct consequence of Lars Hörmander's celebrated regularity theorem [20]. Let $A_1, \dots, A_k$ and $B$ be $C^\infty$ vector fields on $\mathbb{R}^N$, identified with derivation operators, and let $L = -\sum A_j^2 + B$. If the rank of $(A_1, \dots, A_k)$ is strictly less than $N$, then the operator $L$ is not elliptic, and there is no a priori reason why the semigroup $e^{-tL}$ would be regularizing in all variables. But if $-\sum [A_j, B]^2 - \sum A_j^2$ is elliptic, where $[A_j, B]$ is the Lie bracket between $A_j$ and $B$, then $e^{-tL}$ is regularizing in all variables, and the operator $L$ is said to be hypoelliptic. (This is not the classical definition of hypoellipticity, but it will do for the purpose of this presentation.) We see here how the "nondissipative" first-order operator $B$ interacts with the "dissipative part" of $L$, or more precisely the derivation operators $A_j$, to produce the missing directions of regularization. Possibly the most important instance of application is to the operator $L = -\Delta_v + v \cdot \nabla_x$, where $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$; in that case $A_j = \partial/\partial v_j$, $B = v \cdot \nabla_x$, $[A_j, B] = \partial/\partial x_j$. The corresponding evolution equation $\partial_t f + Lf = 0$ is degenerate, but still presents some of the typical features of a parabolic equation; the word "ultraparabolic" is sometimes used for it.

Hypoelliptic regularity has been the object of hundreds of works for the past four decades. But what was understood only very recently is that quite similar phenomena arise in the study of rates of convergence to equilibrium. To describe this, I shall use the word "hypocoercivity", which was suggested to me by Thierry Gallay. A typical hypocoercivity theorem will give sufficient conditions on an operator $L$ so that $e^{-tL}$ will converge to equilibrium at a certain rate, even though $L$ is not "coercive", in the sense that the kernel of its dissipative part is much larger than the set of equilibria.

Hypoellipticity and hypocoercivity are often found together, and have been actually studied together, by refined hypoelliptic techniques [6], [7], [15], [16], [19], and sometimes by probabilistic methods [8], [22], [23]. However, these two phenomena are distinct: Each of them can occur without the other; and the structures which underlie them are not exactly the same. This motivates the development of a separate theory of hypocoercivity. In the sequel, I shall present some of the first results in this direction.

**Acknowledgement.** The ideas exposed in the sequel have benefited from interactions with many people who are quoted within the text. Warm thanks are due to Martin Hairer, Frédéric Hérau and Clément Mouhot for their detailed comments on a preliminary version of these notes; and to Thierry Gallay for illuminating discussions.

## 1. Motivations

In this section I shall describe some concrete examples which motivate the study of hypocoercivity. All of them come from mathematical physics, and none of them is academic. Of course the list is far from exhaustive.

**The kinetic Fokker–Planck equation.** In stochastic analysis, Fokker–Planck equations are often encountered as equations satisfied by the time-dependent laws of solutions of first-order stochastic differential equations. In "real life" however, equations of motion are not first-order, but second-order. Consider for instance a particle in $\mathbb{R}^n$, following Newton's equations with a potential force $-\nabla V$, a white noise random forcing, and a linear friction with coefficient $\theta = 1$: Then its position $X_t$ at time $t$ satisfies the second-order stochastic differential equation

$$\frac{d^2X_t}{dt^2} = -\nabla V(X_t) + \sqrt{2}\,\frac{dB_t}{dt} - \frac{dX_t}{dt},$$

where $B_t$ is a standard Brownian motion. (Of course, the coefficient $\sqrt{2}$ is just a convenient normalization, and the writing is formal in the sense that $B_t$ is not differentiable.) To write the associated partial differential equation, define $f_t(x, v)$ as the density of the law of $(X_t, \dot{X}_t)$ in $\mathbb{R}^n \times \mathbb{R}^n$. Then $f$ is a solution of

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f - \nabla V(x) \cdot \nabla_v f = \Delta_v f + \nabla_v \cdot (fv), \tag{1}$$

where $\Delta_v$ and $\nabla_v \cdot$ respectively stand for the Laplace and divergence operators in velocity space. Equation (1) is *kinetic* in the sense that it involves not only the position, but also the velocity variable; it is one of the fundamental equations in gas dynamics. It admits many nonlinear variants, among which the Vlasov–Poisson–Fokker–Planck equation, which is accepted as one of the fundamental equations of stellar dynamics.

When $V$ is quadratic, the fundamental solution of (1) is explicit and Gaussian. Its examination shows that there is relaxation to a Gaussian equilibrium (in $x$ and $v$ variables) as $t \to \infty$, and this convergence is exponentially fast, with an explicit rate. Here we see a perfect illustration of the hypocoercivity phenomenon: The differential operator on the left-hand side of (1) is conservative (it describes the trajectories of a classical dynamical system in $\mathbb{R}^n \times \mathbb{R}^n$ with Hamiltonian $V(x) + |v|^2/2$), and the right-hand side alone is diffusive degenerate (it only acts on the velocity variable $v$, so cannot cause any relaxation to equilibrium with respect to the $x$ dependence); however, their combination leads to an exponential convergence to equilibrium.

For more general potentials, there is still a global equilibrium:

$$f_\infty(x, v) = \frac{e^{-\left[V(x) + \frac{|v|^2}{2}\right]}}{Z},$$

where $Z$ is a normalizing constant. Then it is an obviously natural question whether exponential convergence to $f_\infty$ holds true under adequate assumptions on the potential $V$, which go beyond the "trivial" quadratic case. Shockingly enough, the first such results were obtained only around 2002, by Frédéric Hérau and Francis Nier [19]. They used a quite sophisticated approach taking roots in Joseph Kohn's approach to hypoellipticity. Since then, their method has been very much simplified, as I shall describe later.

**About the choice of functional space.** The choice of functional space in which to study the large-time behavior of (1) is not innocent. From a probabilistic or physical point of view, it is most natural to assume that $f$ is an integrable density (or even a measure) with possibly rapid decay at infinity. However, in the majority of mathematical studies on the Fokker–Planck equation (kinetic or not), a different choice is made, namely

$$\int \frac{f^2}{f_\infty} < \infty. \tag{2}$$

The reason is simple: Perform a change of unknown in (1) by writing $h = f/f_\infty$ (from a probabilistic perspective, this amounts to considering the adjoint equation); then (1) turns into

$$\frac{\partial h}{\partial t} + v \cdot \nabla_x h - \nabla V(x) \cdot \nabla_v h = \Delta_v h - v \cdot \nabla_v h. \tag{3}$$

Now the operator appearing on the right-hand side is self-adjoint in the Hilbert space $L^2(f_\infty\, dx\, dv)$, so (3) might lend itself to a spectral treatment. Assumption (2) simply says that $h$ belongs to the above-mentioned Hilbert space.

Of course, formally, equations (1) and (3) are equivalent. But this is misleading, since the additional assumption (2) is a very strong restriction. In fact, it does happen that for certain potentials $V$, the convergence to equilibrium is exponential under the "$L^2$-type" assumption (2), but not under a more general "$L^1$-type" assumption (that is assuming just integrability, and maybe some moment bounds). In this sense, $L^1$ results are stronger than $L^2$ results. This is actually one of the reasons of the popularity of logarithmic Sobolev inequalities: They provide a natural functional tool to study convergence to equilibrium in $L^1$ spaces.

The moral of this discussion is that for physical relevance the discussion of convergence to equilibrium of solutions to (1) should not be limited to an $L^2$ framework, but also include more general $L^1$-type assumptions. In the sequel, I shall describe some results in this direction.

**Oscillator chains.** Even though Fourier's law of conduction of heat is one of the oldest partial differential equations, it is still extremely far from a rigorous theoretical understanding. Many models of statistical physics have been proposed to describe heat conduction. Here is one of them, described in [8]. Each atom in a solid body is labelled (for the sake of this discussion, we may assume that the dimension is 1, so atoms are labelled $0, 1, \dots, N$), and the unknowns are the displacements $X^0, \dots, X^N$ of the atoms with respect to their respective equilibrium positions. Each atom is bound to its equilibrium position with a "pinning potential" $V$, and it also interacts with its two neighbors by an interaction potential $W$, assumed to be symmetric ($W(z) = W(-z)$). So the equation for $X^k$ is just

$$\frac{d^2 X_t^k}{dt^2} = -\nabla V(X_t^k) - \nabla W(X_t^k - X_t^{k-1}) - \nabla W(X_t^k - X_t^{k+1}). \tag{4}$$

Of course these equations do not apply to the atoms that are at the extreme left ($k = 0$) and the extreme right ($k = N$) in the chain, since they have only one neighbor. But these extremal atoms are also *shaken* by some external bath, with a temperature of agitation $T^{(\ell)}$ on the left, and $T^{(r)}$ on the right. The corresponding equations for, say, $k = 0$, can be written

$$\begin{cases} \dfrac{d^2 X_t^0}{dt^2} = -\nabla V(X_t^0) - \nabla W(X_t^0 - X_t^1) + \ell, \\ \dfrac{d\ell}{dt} = \lambda^{(\ell)} \sqrt{2T^{(\ell)}}\, \dfrac{dB_t^{(\ell)}}{dt} - \ell + (\lambda^{(\ell)})^2 \ell\, \dfrac{dX_t^0}{dt}. \end{cases} \tag{5}$$

Here $\lambda^{(\ell)}$ is a coefficient describing the strength of the coupling between the particle and the heat bath.

Again, the law of this system is described by a linear partial differential equation in the variables $\ell, r, X^0, \dots, X^N, \dot{X}^0, \dots, \dot{X}^N$. It is very similar to the kinetic Fokker–Planck equation, except that it is much more degenerate, since the diffusion only acts on the variables $\ell$ and $r$.

There are now two difficult problems which naturally arise: (i) Show that the solution $f_t(\ell, r, x^0, \dots, x^N, v^0, \dots, v^N)$ approaches some stationary distribution as $t \to \infty$; (ii) Study the properties of this stationary distribution, and in particular the associated energy flux. (In this case, it is better to say "stationary distribution" rather than "equilibrium", precisely because the temperatures are not necessarily equal.) In particular, if $T^{(\ell)} > T^{(r)}$, in the asymptotic regime $N \to \infty$, is it true that energy flows from the left to the right, and what is the relation between the average flux and the difference of temperatures!?

When $T^{(\ell)} = T^{(r)}$, the equilibrium distribution is easy to write down explicitly, and problem (ii) is trivially solved. But as soon as these temperatures are different, the stationary solution is not explicit – except in the case when $V$ and $W$ are quadratic, but then the results are physically irrelevant!! It is conjectured that some anharmonicity is *necessary* to get the Fourier law (ironically enough, the heat equation, although one of the most basic *linear* models in science, needs some dose of microscopic nonlinearity to be explained). Then problem (ii) becomes incredibly difficult.

Even when the two temperatures are equal, problem (i) appears to be quite difficult. It is actually a typical hypocoercive situation: The diffusion on $\ell$ and $r$ should lead in the end to a relaxation to equilibrium in all variables.

Exponential convergence to the stationary distribution has been proved recently by several authors [8], [7], even for the case when $T^{(\ell)} \neq T^{(r)}$, under various assumptions on the potentials; but the dependence of the estimates upon the number of atoms is just terrible.

**The Boltzmann equation.** The Boltzmann equation is one of the basic partial differential equations in statistical mechanics. It is a kinetic model for the evolution of a rarefied gas of particles interacting via binary collisions. Historically, it has preceded

the Fokker–Planck equation; but the analytical problems that it raises are considerably more acute. A mathematically-oriented presentation of the Boltzmann equation can be found in my long review paper [27]. The classical Boltzmann equation in $n$ dimensions of space can be written

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q(f, f), \quad x \in \Omega_x, \ v \in \mathbb{R}^n, \tag{6}$$

where $\Omega_x$ is a bounded connected open spatial domain, and $Q$ is the Boltzmann collision operator, defined by

$$Q(f, f) = \int_{\mathbb{R}^n} \int_{S^{n-1}} B(v - v_*, \sigma)\big[f(x, v')f(x, v'_*) - f(x, v)f(x, v_*)\big]\, d\sigma\, dv_*.$$

Here $\sigma$ is a variable unit vector in $\mathbb{R}^n$, $B(v - v_*, \sigma)$ is a collision kernel depending on the particular form of the interaction (for instance $B(v - v_*, \sigma) = |v - v_*|$), and the transform $(v, v_*) \to (v', v'_*)$ is computed by the rules of elastic collision:

$$v' = \frac{v + v_*}{2} + \frac{|v - v_*|}{2}\sigma, \quad v'_* = \frac{v + v_*}{2} - \frac{|v - v_*|}{2}\sigma.$$

This equation should of course be supplemented with boundary conditions. To simplify things, one can assume that $\Omega_x$ is just the $n$-dimensional torus $\mathbb{T}^n$ (periodic boundary conditions); another common choice is specular reflexion in a bounded open set.

In spite of hundreds of papers, the mathematical theory of the Boltzmann equation is far from complete; in particular there is still no theory of classical solutions in the large. However, strong regularity results have been obtained in a close-to-equilibrium regime. A complete theory can also be put together as long as there is a pointwise control of certain hydrodynamic fields (density in physical space, mean velocity, temperature, pressure tensor).

It was Boltzmann's beautiful observation that the $H$ functional (negative of the entropy),

$$H(f) = \int f \log f \, dx\, dv$$

is nonincreasing with time along solutions of the Boltzmann equation. Then there is a unique large-time equilibrium, which takes the form of a Maxwellian (Gaussian) distribution:

$$f_\infty(x, v) = \rho \, \frac{e^{-\frac{|v-u|^2}{2T}}}{(2\pi T)^{n/2}},$$

where $\rho \geq 0$ (total mass), $u \in \mathbb{R}^n$ (total mean momentum) and $T \geq 0$ (mean temperature) are constants. This equilibrium is obtained by maximizing the entropy given the conservation laws.

The problem of convergence to equilibrium for the Boltzmann equation is famous for historical reasons (it triggered a hot controversy in the nineteenth century) and also

for theoretical reasons (as a manifestation of irreversibility in the statistical description of a reversible mechanical system; and as the justification of the law of maximum entropy on a basic model). See my lecture notes [29] for an overview of this question.

Of course the complexity of the Boltzmann equation, and its nonlinearity are major difficulties in the study. But behind that, we can recognize once again a hypocoercive situation: The dissipation (collision) operator $Q$ on the right-hand side of (6) is very degenerate since is only acts on the velocity dependence, and it is only its association with the conservative transport operator $v \cdot \nabla_x$ on the left-hand side which can lead to convergence to equilibrium.

**Stability of Oseen's vortices.** The last example in this gallery comes from hydrodynamics and was brought to my attention by Thierry Gallay. It is a well-documented fact in turbulence theory that the vorticity of a two-dimensional incompressible flow tends to coalesce and form large vortices. Thierry Gallay and Eugene Wayne [11] have studied this phenomenon rigorously for a two-dimensional incompressible viscous fluid in the whole space: If $\omega = \omega_t(x)$ is the vorticity, the equation is just

$$\frac{\partial \omega}{\partial t} + \mathrm{BS}[\omega] \cdot \nabla \omega = \Delta \omega,$$

where $\mathrm{BS}[\omega]$ is the velocity field obtained from the vorticity $\omega$ via the Biot–Savart law:

$$\mathrm{BS}[\omega](x) = \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2}\, \omega(y)\, dy,$$

and $v^\perp$ is obtained from $v$ by rotation of angle $\pi/2$.

If $\omega_0 \in L^1(\mathbb{R}^2)$, then $\omega_t$ converges to 0 as $t \to \infty$, due to viscous dissipation. But a refined analysis shows that $\omega_t$ is asymptotically close to an explicit self-similar Gaussian solution, which physically corresponds to a unique large vortex, called Oseen's vortex. In fact, in suitably rescaled variables, the vorticity does converge to a stationary Gaussian distribution.

The linear stability analysis of this phenomenon reduces to the spectral analysis of the operator $S + \alpha B$ in $L^2(\mathbb{R}^2)$, where

$$\begin{cases} S\omega = -\Delta\omega + \dfrac{|x|^2}{16}\,\omega - \dfrac{\omega}{2}, \\ B\omega = \mathrm{BS}[G] \cdot \nabla\omega + 2\,\mathrm{BS}[G^{1/2}\omega] \cdot \nabla G^{1/2}. \end{cases} \tag{7}$$

Here $G$ is a Gaussian distribution: $G(x) = e^{-|x|^2/4}/(4\pi)$; and $\alpha$ is the value of the "circulation Reynolds number", which in the present set of conventions is just $\int \omega_0$.

The spectral study of $S + \alpha B$ turns out to be quite tricky. In the hope of getting a better understanding, one can decompose $\omega$ in Fourier series: $\omega = \sum_{n\in\mathbb{Z}} \omega_n(r) e^{in\theta}$, where $(r, \theta)$ are standard polar coordinates in $\mathbb{R}^2$. For each $n$, the operators $S$ and $B$

can be restricted to the vector space generated by $e^{in\theta}$, and can be seen as just operators on a function $\omega(r)$:

$$\begin{cases} (S_n\omega)(r) = -\partial_r^2\omega - \left(\frac{r}{2} + \frac{1}{r}\right)\partial_r\omega - \left(1 - \frac{n^2}{r^2}\right)\omega, \\ (B_n\omega)(r) = i\,n\,(\varphi\omega - g\Omega_n). \end{cases}$$

Here $g(r) = e^{-r^2/4}/4\pi$, $\varphi(r) = (1-e^{-r^2/4})/2\pi r^2$, and $\Omega_n(r)$ solves the differential equation

$$-(r\Omega')' + \frac{n^2}{r}\Omega = \frac{r}{2}\omega.$$

The regime $|\alpha| \to \infty$ is of physical interest and has already been the object of numerical investigations by physicists. There are two families of eigenvalues which are imposed by symmetry reasons; but apart from that, it seems that all eigenvalues converge to infinity as $|\alpha| \to \infty$, and for some of them the precise asymptotic rate of divergence $O(|\alpha|^{1/2})$ has been established by numerical evidence. If that is correct, this means that the "perturbation" of $S$ by $\alpha B$ is strong enough to send most eigenvalues to infinity as $|\alpha| \to \infty$. This is particularly striking when one realizes that $S$ is symmetric in $L^2(\mathbb{R}^2)$, while $B$ is antisymmetric. Obviously, this is again a manifestation of a hypocoercive phenomenon.

Let us simplify things just a bit by throwing away the nonlocal term $g\Omega_n$ in the expression of $B_n$. After a few manipulations, the problem reduces to the following

**Model Problem 1.1.** Identify sufficient conditions on $f : \mathbb{R} \to \mathbb{R}$, so that the real parts of the eigenvalues of

$$L_\alpha : \omega \longmapsto (-\partial_x^2\omega + x^2\omega - \omega) + i\alpha f\omega$$

in $L^2(\mathbb{R})$ go to infinity as $|\alpha| \to \infty$, and estimate this rate.

So far this problem has been solved only partially, by Isabelle Gallagher and Thierry Gallay; I shall describe their results later on.

## 2. A dynamical approach

Together with Laurent Desvillettes [3], [5], I have developed a method to study quite general hypocoercive situations. The method ultimately relies on the analysis of a *system* of coupled differential inequalities of first and second order (instead of just one first-order differential inequality as in Gronwall's lemma). The method was devised with the aim of proving convergence to equilibrium for *uniformly smooth* solutions of the Boltzmann equation, so I shall explain its principle on that particular example. Complete proofs [5] are quite long and technical, so my goal here is only to isolate the main ideas in a sketchy way.

**First and second order differential inequalities.** As we know, Boltzmann's $H$ functional goes down in time along solutions of (6). A more precise analysis shows that the total entropy production, $-dH/dt$, is always strictly positive, unless $f_t$ is a *hydrodynamical state*:

$$f_t(x, v) = \rho_t(x) \frac{e^{-\frac{|v-u_t(x)|^2}{2T_t(x)}}}{(2\pi T_t(x))^{n/2}}.$$

In words, a hydrodynamical state is a kinetic distribution which is in Maxwellian equilibrium with respect to the velocity variable, but not with respect to the position variable; so it only depends on the fields of local density ($\rho$), mean velocity ($u$) and local temperature ($T$).

With a much more refined analysis, one can establish a *quantitative lower bound* on the entropy production, under adequate (very strong) smoothness, decay and positivity assumptions on $f$: For any $\varepsilon > 0$ there is a constant $K_\varepsilon > 0$ such that

$$-\frac{d}{dt}[H(f) - H(M)] \geq K_\varepsilon \big[H(f) - H(M^f_{\rho\, u\, T})\big]^{1+\varepsilon}, \tag{8}$$

where $M$ is the global equilibrium, which is a Maxwellian distribution, $M^f_{\rho\, u\, T}$ is the hydrodynamical state with the same (local) density, mean velocity and temperature as $f$, and the dependence of $f$, $\rho$, $u$ and $T$ on time is implicit. (The quantity in the right-hand side of (8) is nonnegative.)

Inequality (8) gives a good lower bound on the entropy production, as long as the unknown $f$ *stays away from hydrodynamical states*. But if $f$ decides to become hydrodynamic, or very close to, then the entropy production vanishes and there is nothing that we can deduce about the convergence to equilibrium. This is where the antisymmetric part of the Boltzmann equation has to help us.

Now the second differential inequality is obtained by introducing a suitable functional measuring the distance of $f$ to the space of hydrodynamical states. On one hand it should be controlled by the quantity $H(f) - H(M^f_{\rho\, u\, T})$; but on the other hand it should be simple enough to make explicit computations. A natural choice is $\mathcal{E}_1(f) = \|f - M^f_{\rho\, u\, T}\|^2_{L^2}$; note that this functional depends on $f$ in a strongly nonlinear way, via $\rho$, $u$ and $T$. The point now is to give a *lower bound on the second-order time-derivative* of this new functional.

Differentiating a functional once along the Boltzmann equation is already complicated, but differentiating it twice is a horrendous task; so it better be well motivated. There are in fact two main reasons to consider the second derivative. The first is that the second derivative in time gives a measure of how fast the distance between $f$ and $M^f_{\rho\, u\, T}$ will increase again if it ever vanishes, or becomes very close to. The second reason is that by applying twice the Boltzmann equation, we let the first-order operator $v \cdot \nabla_x$ act twice, and then the resulting computations are somewhat similar to those that would have been obtained by letting a second-order operator *in the $x$ variable* act once. So in this second-order time derivative one will find some of the

terms that would have appeared in a first-order computation along the heat equation in the $x$ variable.

So, after many calculations one can show that for $\delta_1$ small enough,

$$\frac{d^2}{dt^2}\|f - M^f_{\rho\, u\, T}\|^2_{L^2} \geq K_1 \left[\int_{\Omega_x} |\nabla T(x)|^2\, dx + \int_{\Omega_x} |\{\nabla u(x)\}|^2\, dx\right] - \frac{C_1}{{\delta_1}^{1-\varepsilon}}\left(\|f - M^f_{\rho\, u\, T}\|^2_{L^2}\right)^{1-\varepsilon} - \delta_1[H(f) - H(M)], \tag{9}$$

where $K_1$ and $C_1$ are constants which only depend on some a priori smoothness and positivity estimates on $f$, and $\{\nabla u\}$ is the traceless part of the symmetric part of the matrix-valued field $\nabla u$.

To understand what has been achieved, assume for a moment that $f$ becomes hydrodynamical at some time $t_0$, and forget the error term with $\delta_1$. Then we have $(d^2/dt^2)\|f - M^f_{\rho\, u\, T}\|^2_{L^2} \geq K_1\|\nabla T\|^2_{L^2}$. In particular, $\|f - M^f_{\rho\, u\, T}\|^2$ is strictly convex, as a function of $t$, unless $\nabla T$ is equal to 0, and will grow quadratically for a short time. So this second equation gives us some information about the inhomogeneities in the temperature field $T$. In a geometric language, what we have shown, more or less, is that the Boltzmann flow is "transverse" to the space of hydrodynamical states, in presence of heterogeneities of the temperature.

Combining this with the first inequality (8), we see that we have some information about how far $f$ is to the space of hydrodynamical states with constant temperature. There is still something missing, but now we can repeat the procedure: Introduce a new functional measuring the distance of $f$ to that space, for instance $\mathcal{E}_2(f) = \|f - M^f_{\rho\, u\, \langle T\rangle}\|^2_{L^2}$, where $M^f_{\rho\, u\, \langle T\rangle}$ is the hydrodynamical state which has the same density $\rho$ and velocity fields $u$ as $f$, and a constant temperature whose value is computed by averaging the temperature $T$ of $f$ against the density $\rho$. Then differentiate twice again. New computations yield a result which is very similar to the one in (9), except that the terms $\|\nabla T\|^2$ and $\|\{\nabla u\}\|^2$ are replaced by $\|\nabla^{\mathrm{sym}} u\|^2$, the square $L^2$ norm of the complete symmetric part of $\nabla u$.

As a general principle (Korn's inequality), a control of the symmetric part of $\nabla u$ implies a control on the whole of $\nabla u$, under suitable boundary conditions. In fact the boundary conditions here are not standard, at least for specular reflexion, so the desired estimates do not follow from the classical theory of Korn inequalities; but let us forget this for the moment and assume that we have indeed a good control on the inhomogeneities of the velocity field.

So far there is still no control on the inhomogeneities of the density, but at this stage the reader has probably understood the sequel of the method: Introduce a suitable functional $\mathcal{E}_3$ measuring the distance of $f$ to the space of hydrodynamical states with constant temperature and constant velocity field, and differentiate this expression twice in time. It is possible to choose $\mathcal{E}_3(f) = \|f - M^f_{\rho\, 0\, 1}\|^2_{L^2}$, where $M^f_{\rho\, 0\, 1}$ has the same density as $f$, the same average temperature and the same total momentum; and then in the resulting computations pops up the desired term $\|\nabla\rho\|^2$.

**Closing the system.** At this stage we have a system of four differential inequalities (one of first order, and three of second order). To close the system, one can use:

- A *physical input*: The total entropy is the sum of a purely kinetic entropy and a purely hydrodynamical entropy, which controls inhomogeneities of all fields $\rho$, $u$ and $T$.
- An *analytical input*: In presence of smoothness bounds, all the norms appearing in the computations are "almost equivalent". More precisely, if $\|f\|_1$ and $\|f\|_2$ are any two Lebesgue, or Sobolev norms, then for any $\varepsilon > 0$ one can find a constant $C_\varepsilon$, only depending on some smoothness estimates on $f$, such that $\|f\|_1 \leq C_\varepsilon \|f\|_2^{1-\varepsilon}$. This step, obviously based on elementary interpolation theory, is crucial to "get the exponents right"; without it, one would get disastrous rates of convergence, or just no rate at all. This way of *trading smoothness for exponents* is one of the reasons why the method is so greedy in regularity.
- A *geometric/analytical input*: Certain norms of differential quantities ($\|\nabla T\|$, $\|\nabla^{\mathrm{sym}} u\|$, $\|\nabla \rho\|$) imply a control on the departure of the corresponding fields to their mean value. The shape of the domain (connectedness, rotational symmetry, etc.) plays a crucial role here; all of this can be expressed *quantitatively* with some functional inequalities of Poincaré or Korn type (see in particular [4]).

**Study of the differential system.** Not all the hard job has been done at this stage. The result is a system of first and second-order differential inequalities, coupled together in a quite intricate way, from which one wants to extract estimates about the rates of relaxation. This can be done with the help of the following (definitely not obvious) lemma:

**Lemma 2.1.** *Let $h(t) \geq 0$ satisfy*

$$h''(t) + A h(t)^{1-\varepsilon} \geq \alpha > 0 \quad \text{for all } t \in (t_1, t_2)$$

*for some $\varepsilon < 0.1$. Then,*

*– either $t_2 - t_1$ is small:*

$$t_2 - t_1 \leq 50\, \frac{\alpha^{\frac{\varepsilon}{2(1-\varepsilon)}}}{A^{\frac{1}{2(1-\varepsilon)}}};$$

*– or $h$ is large on the average:*

$$\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} h(t)\, dt \geq \frac{\alpha^{\frac{1}{1-\varepsilon}}}{100}\ \inf\left(\frac{1}{A}, \frac{1}{A^2}\right).$$

With a repeated use of Lemma 2.1, one can show in the end that the system of differential inequalities implies relaxation to equilibrium at a rate $O(t^{-\infty})$, that is, faster than any inverse power of time. All in all, one can get the following result, which is stated here in a slightly sketchy way (see [5] for more precise statements):

**Theorem 2.2.** *Let $(f_t)_{t\geq 0}$ be a smooth solution of the Boltzmann equation* (6)*, such that all the derivatives of $f$ are uniformly bounded, and all the moments of $f$ are bounded, uniformly in time. Further assume that $f$ satisfies a pointwise lower bound of the form $f_t(x,v) \geq K_0 e^{-A_0|v|^{q_0}}$. Then, under adequate boundary conditions, $f_t$ converges to global equilibrium as $t \to \infty$, at least as fast as $O(t^{-\kappa})$ for all $\kappa > 0$.*

**Further comments.** More information about the implementation of this program in the context of the Boltzmann equation can be found in the original research paper [5], or, in a lighter form, in the lecture notes [30], [29]. Before being used on the Boltzmann equation, the dynamical approach had been tried on the Fokker–Planck equation [3] and on some other linear models [1], [9]. It is quite robust and adapted to equations with very little structure.

One of its appealing features is that it seems to provide a good physical intuition of what is going on: The system approaches hydrodynamical state under the influence of collisions, then it is driven out of hydrodynamical state by the influence of the transport, etc. Numerical simulations have corroborated this qualitative analysis surprisingly well. In the diagram below, computed numerically by Francis Filbet, one sees very clearly that the solution of the Boltzmann equation oscillates between states

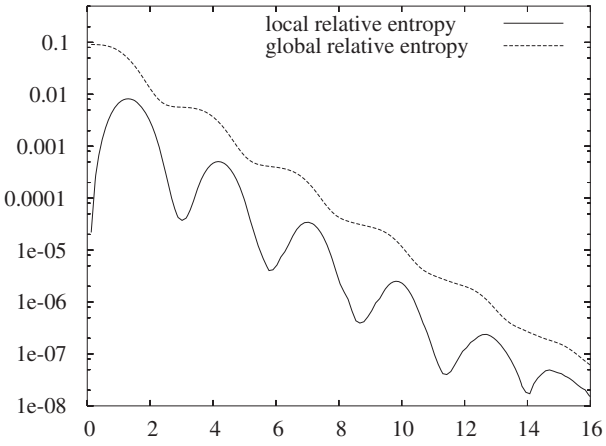



Figure 1. The upper curve is the $H$ functional as a function of time, in semi-log plot; the lower curve is the purely kinetic part of the $H$ functional. When the two curves are far away, the distribution is almost in hydrodynamical state; when they are very close, it is almost homogeneous.

where it is close to hydrodynamical, and states where it is close to homogeneous. In particular, contrary to a widespread belief, the approach to hydrodynamical regime is not faster than the approach to global equilibrium. (All of this is valid only on scales of time on which the Knudsen number is of order 1.)

From the point of view of physics, the discovery of these oscillations may be one of the most noticeable outcomes of the program of hypocoercivity applied to

the Boltzmann equation. (No doubt that one day there will be a simpler analytical way to explain them.) They are not easy to observe, and have even been used as a "benchmark" to test the accuracy of certain numerical schemes (see e.g. [10]).

However, the dynamical method suffers from the complexity of its practical application, and its heavy computational cost. In the next section, I shall describe another method which may be less appealing from the physical point of view, and requires a bit more structure, but has the advantage to be much lighter.

Also, I emphasize that Theorem 2.2 requires strong regularity and decay estimates on the solutions. As discussed in [5], all these estimates can be proven in the *close-to-equilibrium* regime, but remain a major open problem for solutions in the large. Even in a close-to-equilibrium regime, decay estimates based on a linearization method are quite hard to obtain, and were not available at the time when [5] was published; since then this gap has been filled in a series of important works by Yan Guo and Robert Strain [13], [14].

## 3. A functional approach

In the previous method the resolution of the main degeneracy problem was done via the time-differentiation of certain (relatively) simple functionals. Now the idea is to put as much as possible of the difficulty in a careful choice of the functional; and more precisely to add "correction terms" which are negligible in size, but contribute in an important way to the time-derivative of the functional. This will be more easily explained on the example of the Fokker–Planck equation, in the form (3).

Suppose you want to get a Gronwall inequality for some well-chosen functional, applied to the Fokker–Planck equation. First try the $L^2$ norm. With the notation $\mu(dx\,dv) = f_\infty(x,v)\,dx\,dv$, and omitting once again the dependence upon time, one has

$$\frac{d}{dt}\int h^2\,d\mu = -\int |\nabla_v h|^2\,d\mu.$$

Since the derivatives in the right-hand side involve only the velocity variables, there is no way to dominate the integral in the left-hand side by the right-hand side (choose $h = h(x)$, then the right-hand side vanishes).

So go to a higher order norm, involving gradients of $h$. After a bit of work, under suitable assumptions on $V$, one can find constants $a, c, K > 0$ such that

$$\begin{aligned}\frac{d}{dt}\left(\int h^2\,d\mu + a\int |\nabla_x h|^2\,d\mu + c\int |\nabla_v h|^2\,d\mu\right)\\ \leq -K\left(\int |\nabla_v h|^2\,d\mu + \int |\nabla_v\nabla_x h|^2\,d\mu + \int |\nabla_v\nabla_v h|^2\,d\mu\right).\end{aligned} \tag{10}$$

Again, the right-hand side is not sufficient to control the expression in brackets on the left-hand side. So still nothing!

But now correct the functional on the left-hand side by adding an innocent-looking term $2b \int \nabla_x h \cdot \nabla_v h \, d\mu$. If $b < \sqrt{ac}$, this term does not play any noticeable role in the value of the functional, since

$$\left| 2b \int \nabla_x h \cdot \nabla_v h \, d\mu \right| \leq (1 - \delta) \left[ a \int |\nabla_x h|^2 \, d\mu + c \int |\nabla_v h|^2 \, d\mu \right]$$

for some positive constant $\delta$. However, if $a$, $b$ and $c$ are properly chosen, then we have a differential inequality which is much better than (10):

$$\begin{aligned} &\frac{d}{dt} \left( \int h^2 \, d\mu + a \int |\nabla_x h|^2 \, d\mu + 2b \int \nabla_x h \cdot \nabla_v h \, d\mu + c \int |\nabla_v h|^2 \, d\mu \right) \qquad (11) \\ &\leq -K \left( \int |\nabla_x h|^2 \, d\mu + \int |\nabla_v h|^2 \, d\mu + \int |\nabla_v \nabla_x h|^2 \, d\mu + \int |\nabla_v \nabla_v h|^2 \, d\mu \right). \end{aligned}$$

Now it very easy to close this differential inequality: It suffices that $\mu$ satisfies a Poincaré inequality (in the $x$ and $v$ variables).

**The algebraic core.** I started to work on this approach while struggling to understand the results of Frédéric Hérau and Francis Nier [19], without resorting to the technical hypoelliptic machinery used in their work. After deciding that there should be an elementary approach based on integration by parts and chain rule, I was still flooded by the complex calculations. Then I decided that there should be an even simpler approach with no analysis at all. After going to an abstract formulation of the problem, I found out that there was indeed an extremely simple "algebraic core" which can be presented as follows. Take two operators $A$ and $B$ on a Hilbert space (in the present case $A$ would be the vector-valued differential operator $\nabla_v$, while $B$ would be $v \cdot \nabla_x - \nabla V(x) \cdot \nabla_v$), with $B^* = -B$. Then, at least formally, the time-derivative of

$$\langle Ah, [A, B]h \rangle$$

along the influence of $B$ can be written as

$$\langle ABh, [A, B]h \rangle + \langle Ah, [A, B]Bh \rangle.$$

Pretend that $B$ commutes with $[A, B]$; then the previous expression is

$$\langle ABh, [A, B]h \rangle + \langle Ah, B[A, B]h \rangle,$$

and since $B^* = -B$, this can be rewritten as

$$\langle ABh, [A, B]h \rangle - \langle BAh, [A, B]h \rangle = \langle [A, B]h, [A, B]h \rangle. \qquad (12)$$

In the example of the Fokker–Planck equation, $[A, B] = \nabla_x$, so $\langle Ah, [A, B]h \rangle = \int \nabla_v h \cdot \nabla_x h \, d\mu$, and the right-hand side of (12) is the desired term in $\|\nabla_x h\|^2$.

The advantage to input terms with "mixed derivatives" such as $\int \nabla_x h \cdot \nabla_v h$ had been actually noted before in studies of global in time propagation of the smoothness

for kinetic equations, most notably by Denis Talay [24] and Yan Guo [12]. The simple algebraic core presented above explains why this trick also applies to problems of convergence to equilibrium.

The rest of this section will be devoted to a presentation of some results which have been obtained by pushing further this approach. All the results quoted below are extracted from two preprints by the author [31], [32], and another preprint by Clément Mouhot and Lukas Neumann [21]. There is also an independent series of works by Frédéric Hérau [18], [17], which is based on quite similar tools.

**The basic theorem.** Two important features of the next theorem are that

– it applies to a general abstract framework: $\mathcal{H}$ is a Hilbert space (think of $\mathcal{H}$ as $L^2(\mu)$, where $\mu$ is the equilibrium measure); and $\mathcal{V}$ another Hilbert space (think of $\mathcal{V}$ as $\mathbb{R}^n$, the space of velocities); then $A$ is an unbounded operator $\mathcal{H} \to \mathcal{H} \otimes \mathcal{V}$, and $B$ is an unbounded operator $B \to B$ with $B^* = -B$;

– it considers a linear operator $L$ which is in (abstract) *Hörmander form*, that is $L = A^*A + B$ for some operators $A$, $B$ as above.

Assume that the semigroup $e^{-tL}$ is well defined and that there is no problem to differentiate the square norms, etc. Systematic tensorization with the identity operator will be used to make sense of notation such as $[A, B] = AB - (B \otimes I)A$. The scalar product in $\mathcal{H}$ will be denoted by $\langle \cdot, \cdot \rangle$, the norm in $\mathcal{H}$ by $\| \cdot \|$, and an operator $S$ will be said to be bounded respectively to a family of operators $T_1, \dots, T_k$ if there is a constant $C$ such that $\|Sy\| \le C(\|T_1 y\| + \cdots + \|T_k y\|)$. The symbol $\Re$ stands for real part.

**Theorem 3.1.** *With the above notation, write* $[A, B] = C$*, and assume that*

(i) $[A, C] = 0$, $[A^*, C] = 0$;

(ii) $[A, A^*]$ *is bounded relatively to* $I$ *and* $A$;

(iii) $[B, C]$ *is bounded relatively to* $A$, $A^2$, $C$ *and* $AC$.

*Further assume that*

(H) $A^*A + C^*C$ *is coercive.*

*Then for a suitable choice of constants* $a, b, c$ *one has the differential inequality*

$$\frac{d}{dt}\mathcal{F}(e^{-tL}h) \le -K\,\mathcal{F}(e^{-tL}h),$$

*where*

$$\mathcal{F}(h) = \|h\|^2 + a\,\|Ah\|^2 + 2b\,\Re\,\langle Ah, Ch\rangle + c\,\|Ch\|^2,$$

*and* $K$ *is a positive constant which only depends on the constants appearing implicitly in assumptions* (ii)*,* (iii) *and* (H).

**Remark 3.2.** The assumption (H) is obviously an analog in this context of Lars Hörmander's bracket condition.

This theorem applies to the Fokker–Planck equation (3) under simple assumptions on the potential $V$, and yields exponential convergence to equilibrium for initial data $h_0$ satisfying $\|\nabla_x h_0\|^2 + \|\nabla_v h_0\|^2 < +\infty$. The latter restriction can finally be removed by an independent study of hypoelliptic regularity [16], [31]. (This is not a standard hypoelliptic estimate since it is global; there would be much to say about it, but this would take us too far.) In the end, one obtains the following theorem, which generalizes and improves the results of [19], [16]. Recall that a measure $\nu$ is said to satisfy a Poincaré inequality if one has a functional inequality of the form $\|\nabla h\|_{L^2(\nu)} \geq P\|h - \langle h\rangle\|_{L^2(\nu)}$, $P > 0$, where $\langle h\rangle$ is the average value of $h$ with respect to $\nu$.

**Theorem 3.3.** *Let $V \in C^2(\mathbb{R}^n)$ with $\inf V > -\infty$, such that*

(a) $|\nabla^2 V| \leq C(1 + |\nabla V|)$;

(b) *the reference measure $\nu(dx) = e^{-V(x)}\,dx$ satisfies a Poincaré inequality with constant $P$.*

*Let $\mu(dx\,dv) = e^{-(V(x)+|v|^2/2)}\,dx\,dv/Z$, where $Z$ is a normalizing constant. Then there are constants $\lambda > 0$ and $C'$, explicitly computable in terms of $C$ and $P$, such that solutions of the Fokker–Planck equation* (3) *satisfies*

$$\|h_t - \langle h_0\rangle\|_{L^2(\mu)} \leq C' e^{-\lambda t}\,\|h_0 - \langle h_0\rangle\|_{L^2(\mu)}.$$

Theorem 3.1, or more precisely its proof, was also used by Isabelle Gallagher and Thierry Gallay to provide a first solution to the Model Problem 1.1, as follows. Set $\mathcal{H} = L^2(\mathbb{R};\mathbb{C})$, $A = \partial_x\omega + x\omega$, $B\omega = (i\alpha f)\omega$. Then $C\omega = i\alpha f'\omega$, so the operator $A^*A + C^*C$ is of Schrödinger type:

$$(A^*A + C^*C)\,\omega = (-\partial_x^2\omega + x^2\omega - \omega) + \alpha^2 f'^2\omega,$$

and the spectrum of $A^*A + C^*C$ can be studied via standard semi-classical techniques. For instance, if $f'(x)^2 = x^2/(1+x^2)^k$, $k \in \mathbb{N}$, then the spectral gap of $A^*A + C^*C$ is bounded below like $O(|\alpha|^{2\nu})$, with $\nu = \min(1, 2/k)$. Then a careful examination of the proof of Theorem 3.1 yields a lower bound like $O(|\alpha|^{\nu})$ on the real part of the spectrum of $A^*A + B$.

**Multiple commutators.** As in Lars Hörmander's hypoellipticity theorem, multiple commutators are also allowed in hypocoercivity results. But as an important difference, it seems that one only needs to consider commutators with the antisymmetric part. Here is such a theorem:

**Theorem 3.4.** *With the same notation as before, assume the existence of (possibly unbounded) operators $C_0, C_1, \ldots, C_{N_c+1}$, $R_1, \ldots, R_{N_c+1}$, and $Z_1, \ldots, Z_{N_c+1}$ such that*

$$C_0 = A, \quad [C_j, B] = Z_{j+1}C_{j+1} + R_{j+1} \ \ (0 \leq j \leq N_c), \quad C_{N_c+1} = 0,$$

*and, for all* $k \in \{0, \dots, N_c\}$,

(i) $[A, C_k]$ *is bounded relatively to* $\{C_j\}_{0\le j\le k}$ *and* $\{C_j A\}_{0\le j\le k-1}$;

(ii) $[A^*, C_k]$ *is bounded relatively to* $I$ *and* $\{C_j\}_{0\le j\le k}$;

(iii) $R_k$ *is bounded relatively to* $\{C_j\}_{0\le j\le k-1}$ *and* $\{C_j A\}_{0\le j\le k-1}$;

(iv) $Z_j$ *is bounded relatively to* $I$, *and* $I$ *is bounded relatively to* $Z_j$;

(H) $\sum_{j=0}^{N_c} C_j^* C_j$ *is coercive.*

*Then one can choose constants* $a_k$ *and* $b_k$ *in such a way that the functional*

$$\mathcal{F}(h) = \|h\|^2 + \sum_{k=0}^{N_c} \big( a_k \|C_k h\|^2 + 2 b_k \, \Re \, \langle C_k h, C_{k+1} h \rangle \big)$$

*satisfies the differential inequality*

$$\frac{d}{dt} \mathcal{F}(e^{-tL} h) \le -K \mathcal{F}(e^{-tL} h)$$

*for some constant* $K$ *which can be computed explicitly in terms of the constants appearing implicitly in* (i)–(iv) *and* (H).

This result generalizes Theorem 3.1 in several ways: Multiple commutators are allowed; a remainder $R_{j+1}$, and a multiplier $Z_j$ are allowed in the identity defining $C_{j+1}$ in terms of $C_j$; and the various directions $C_k$ are not assumed to commute.

As a simple application of Theorem 3.4, it is possible to prove exponential convergence to equilibrium for the oscillator chain described by equations (4)–(5), under the assumption that $V$ and $W$ are uniformly convex and have a bounded Hessian, *and that the temperatures on the left and on the right are equal*, that is $T^{(\ell)} = T^{(r)}$. Interestingly enough, bounded Hessians are not covered by the results of Jean-Pierre Eckmann and Martin Hairer [7], who impose a superquadratic growth at infinity. Conversely, it is not clear whether Theorem 3.4 can be used to recover the results in [7]. Still some work is required to clarify the situation about these assumptions on the potentials. I shall also come back in the end of these notes to the very unsatisfactory restriction $T^{(\ell)} = T^{(r)}$.

**$L \log L$ estimates.** Now it is not so difficult to adapt the previous $L^2$ theory to an $L \log L$ framework, replacing $L^2$ square norms by *entropies and Fisher informations*. For this I shall have to leave the framework of abstract Hilbert spaces, and replace it by functional spaces on, say, $\mathbb{R}^N$ (or a differentiable manifold). Then the operators $A_1, \dots, A_m$ and $B$ will be vector fields, identified with differentiation operators, the notation $A$ will stand for the vector-valued differential operator $A = (A_1, \dots, A_m)$, and I shall now say that $S$ is bounded relatively to $T_1, \dots, T_k$ if there is a constant $C$ such that $|S(x)| \le C(|T_1(x)| + \cdots + |T_k(x)|)$, where $S(x)$ stands for the value of the vector field $S$ at $x$. The equilibrium measure will be assumed to take the form $\mu(dX) = e^{-E} \, dX$, where $E$ is a smooth function, $dX$ is the Lebesgue measure in

$\mathbb{R}^N$, and $E$ is normalized so that $\int e^{-E} = 1$. The notation $S^*$ will stand for the adjoint of $S$ *in* $L^2(\mu)$. The linear equation under study will still be $\partial_t h + Lh = 0$, where $L = A^*A + B$ and the unknown is the probability density $f = he^{-E}$.

**Theorem 3.5.** *With the above conventions, assume that all the assumptions in Theorem* 3.4 *are satisfied, up to the following reinforcements:*

(i′) $[A, C_k]$ *is bounded relatively to* $A$;

(ii′) $[A, C_k^*]$ *and* $[A, C_k]^*$ *are bounded relatively to* $A$ *and* $I$;

(iii′) $R_k$ *is bounded relatively to* $\{C_j\}_{0\le j\le k-1}$.

*Further assume that there is a positive constant* $\lambda$ *such that* $\sum_k C_k^* C_k \ge \lambda I_N$, *pointwise on* $\mathbb{R}^N$, *and that* $\mu$ *satisfies a logarithmic Sobolev inequality. Then there are quadratic forms* $x \to S(x)$, *uniformly positive definite, such that the functional*

$$\mathcal{I}(f) = \int f(\log f + E) + \int f\big\langle S\,\nabla(\log f + E), \nabla(\log f + E)\big\rangle$$

*satisfies*

$$\frac{d}{dt}\,\mathcal{I}\big((e^{-tL}h)e^{-E}\big) \le -K\,\mathcal{I}\big((e^{-tL}h)e^{-E}\big),$$

*for some explicitly computable constant* $K$.

The most noticeable novelty in the assumptions of Theorem 3.5 is that now the reference measure is not required to satisfy a Poincaré inequality, but a logarithmic Sobolev inequality, i.e. for any probability density $f$ one should have

$$\int f\,(\log f + E) \le (2P)^{-1} \int f\,|\nabla(\log f + E)|^2,$$

where $P$ is a positive constant. (The normalization here is such that $e^{-E}$ automatically satisfies a Poincaré inequality with constant $P$.)

Apart from this, Theorem 3.5 looks very similar to Theorem 3.4. In fact, by writing densities in the form $f = (1+\varepsilon h)e^{-E}$ and letting $\varepsilon \to 0$, one can recover the conclusion of Theorem 3.4 as a perturbative limit regime. Still, there are some subtle things going on, as indicated by the reinforced assumptions (i')–(iii'). In fact there are some tricky additional computations underlying the proof, with rather miraculous simplifications, suggesting that an adequate formalism is still to be found.

Here is an application of Theorem 3.5:

**Theorem 3.6.** *Let* $V$ *be a* $C^\infty$ *potential on* $\mathbb{R}^n$; *assume that there are positive constants* $k$, $C$, $\{C_j\}_{j\in\mathbb{N}}$ *such that*

(a) $|\nabla^j V(x)| \le C_j$ *for all* $j \ge 2$;

(b) $e^{-V}$ *satisfies a logarithmic Sobolev inequality.*

*Then for any initial datum* $f_0(x, v)$ *with finite moments of all orders, the solution of the Fokker–Planck equation* (1) *converges to equilibrium exponentially fast in the sense of relative entropy and* $L^1$ *norm, with a rate of exponential convergence that does not depend on* $f_0$.

This result is obtained by combining Theorem 3.5 with a global hypoelliptic regularization theorem for $L^1$ initial data. (This again is a highly nonstandard framework for regularization, but I shall not develop this here.)

**Beyond the Hörmander form.** All the examples treated so far were dealing with linear operators in the form $L = A^*A + B$, where $B$ is antisymmetric. In theory, any operator can of course be cast in this form, but this might be a terrible thing to do in practise; for instance, if the symmetric part of $L$ is an integral operator then $A$ would look horrendous. So it is desirable to prove results under alternative structure assumptions.

It would be illusory to hope for a gain based on commutators like $[L, B]$. Instead, one can introduce an adequate auxiliary operator $A$ into the estimates, in such a way that (i) $A^*A + [A, B]^*[A, B]$ is coercive, and (ii) $A$ "almost commutes" with $L$.

Recently, Clément Mouhot and Lukas Neumann [21] have derived such a hypocoercivity theorem in the particular framework of kinetic equations; more precisely, $A = \nabla_v$, $B = v \cdot \nabla_x$, $C = \nabla_x$, and

$$L = v \cdot \nabla_x - \mathcal{L}, \quad \mathcal{L} = K - \Lambda, \tag{13}$$

where $K$ and $\Lambda$ only act on the velocity variable, $-\Lambda$ is "damping" (for instance a multiplication operator) and $K$ is "regularizing" (for instance an integral operator). These assumptions cover many interesting cases in kinetic theory [21]. Here I state the results in a slightly more precise (although not yet fully rigorous) way:

**Theorem 3.7.** *Let $\mathcal{L}$ be an unbounded operator on $L^2(\mathbb{R}^n_v)$, taking the form $\mathcal{L} = K - \Lambda$, and assume that there exists a Hilbert norm $\|\cdot\|_\Lambda$, with $\|\cdot\|_\Lambda \geq \|\cdot\|_{L^2}$, and constants $\kappa, C > 0$ such that*

(i) $\kappa \, \|h\|_\Lambda^2 \leq \langle \Lambda h, h\rangle_{L^2} \leq C \|h\|_\Lambda^2$;

(ii) $\langle \mathcal{L} h, g\rangle \leq C \|h\|_\Lambda \|g\|_\Lambda$;

(iii) *for all $\delta > 0$ there exists $C_\delta > 0$ such that $\langle \nabla_v K h, \nabla_v h\rangle_{L^2} \leq C_\delta \|h\|_{L^2}^2 + \delta \|\nabla_v h\|_{L^2}^2$ for all $h$*;

(iv) $\langle \nabla_v \Lambda h, \nabla_v h\rangle_{L^2} \geq \kappa \|\nabla_v h\|_{L^2}^2 - C \|h\|_{L^2}^2$;

(C) $\langle \mathcal{L} h, h\rangle_{L^2} \leq -\kappa \, \|h - \Pi h\|_{L^2}^2$, *where $\Pi$ is the orthogonal projection onto the kernel of $\mathcal{L}$, assumed to be finite-dimensional.*

*Then the operator $L = -v \cdot \nabla_x + \mathcal{L}$ is hypocoercive in $L^2(\mathbb{T}^n_x \times \mathbb{R}^n_v)$. There are constants $a, b, c, \lambda > 0$ such that the functional defined by*

$$\mathcal{F}(h) = \|h\|_{L^2}^2 + a \, \|\nabla_v h\|_{L^2}^2 + 2b \, \langle \nabla_v h, \nabla_x h\rangle_{L^2} + c \, \|\nabla_x h\|_{L^2}^2$$

*satisfies*

$$\frac{d}{dt} \mathcal{F}(e^{-tL} h) \leq -\lambda \, \mathcal{F}(e^{-tL} h)$$

*for all $h \in L^2(\mathbb{R}^n_x \times \mathbb{R}^n_v) / \operatorname{Ker} L$.*

Some comments on the assumptions: Assumption (i) implies the damping nature of $-\Lambda$, and assumption (iii) is a very weak way to state the regularizing property of $K$. Assumption (iv) is some way to express the fact that $\nabla_v$ and $\Lambda$ satisfy good commutation relations: The estimate would be trivial if $\nabla_v \Lambda$ were replaced by $\Lambda \nabla_v$. Finally, assumption (C) expresses the coercivity of the operator $\mathcal{L}$ when applied to functions which only depend on the velocity variable.

Theorem 3.7 applies for instance to the linearized Boltzmann equation, or many other linear Boltzmann-type models. (There is also an independent study by Frédéric Hérau [17] which analyzes the hypocoercivity of some such operators with very similar tools.)

By combining Theorem 3.7 with a linearization analysis and some global bounds derived by Yan Guo, Clément Mouhot and Lukas Neumann were able to recover a simple proof of the following theorem of convergence to equilibrium for the *nonlinear* Boltzmann equation. In the next statement, $H^k$ stands for the standard $L^2$-Sobolev space of order $k$ on the domain $\mathbb{T}^n_x \times \mathbb{R}^n_v$.

**Theorem 3.8.** *Consider the Boltzmann equation* (6) *in* $\mathbb{T}^n_x \times \mathbb{R}^n_v$*, with the collision kernel* $|v - v_*|$*. Let* $f_0$ *be a* $C^\infty$ *initial density with associated global equilibrium* $f_\infty(x, v) = M(v) = e^{-|v|^2/2}/(2\pi)^{n/2}$*. If*

$$\|M^{-1/2}(f_0 - M)\|_{H^k} \leq \varepsilon,$$

*for some* $k$ *large enough and some* $\varepsilon > 0$ *small enough, then the corresponding solution of the Boltzmann equation converges to equilibrium exponentially fast:*

$$\|M^{-1/2}(f_t - M)\|_{L^2} = O(e^{-\lambda t}).$$

**Nonlinear equations.** To conclude this section, I shall show how to recover fully nonlinear hypocoercivity estimates by a variant of the approach developed above. For general nonlinear operators, it is probably hopeless to try to get anywhere unless one assumes some strong assumptions of smoothness and decay at infinity, to make sure that all norms involved are "almost comparable" (that is, they are comparable if one allows them to be raised to powers that are arbitrarily close to 1). So I will assume that $(f_t)_{t \geq 0}$ satisfies uniform bounds in a scale of weighted Sobolev spaces $(X^s)_{s \in \mathbb{R}}$ of arbitrarily high smoothness and decay, that are in interpolation. (For instance, $X^s$ might be defined as the space of functions $f$ such that $(I - \Delta_v - \Delta_x)^{s/2} f(x, v)(1 + |x|^2 + |v|^2)^{s/2}$ lies in $L^2$.) Then all the nonlinear operators involved will be assumed to be Lipschitz when restricted on balls of $X^s$, with values in some higher order space $X^{s+k}$. In practise, this means that our nonlinearities are not worse than polynomial, with coefficients that do not increase faster than polynomial. Then I shall denote the functional derivative of a functional $\mathcal{F}$ at function $f$ by just $\mathcal{F}'_f$. I shall further assume that there is a unique equilibrium $f_\infty$, and a Lyapunov functional $\mathcal{E}$ satisfying

$$\mathcal{E}(f_t) - \mathcal{E}(f_\infty) \geq K \, \|f_t - f_\infty\|_s^{2(1+\varepsilon)}$$

for some suitable $s = s(\varepsilon)$, $K = K(\varepsilon)$, where $\varepsilon$ is arbitrarily small. In words, this means that $\mathcal{E}$ essentially controls the square of the distance to equilibrium.

**Theorem 3.9.** *With the above notation, let*

$$L = B - \mathcal{C}$$

*be a nonlinear differential operator, such that $B$ preserves the Lyapunov functional $\mathcal{E}$ (that is, $\mathcal{E}'_f \cdot Bf = 0$), let $(f_t)_{t\geq 0}$ solve $\partial_t f + Lf = 0$, and let $(\Pi_j)_{1\leq j\leq J}$ be nonlinear operators satisfying*

$$\Pi_j \circ \Pi_k = \Pi_{\max(j,k)}, \tag{14}$$

*such that, for all $t \geq 1$,*

(i) $\mathcal{C} \circ \Pi_1 = 0$; $-\mathcal{E}'_{f_t} \cdot (\mathcal{C} f_t) \geq K_\varepsilon [\mathcal{E}(f) - \mathcal{E}(\Pi_1 f)]^{1+\varepsilon}$;

(ii) $K_\varepsilon \|\Pi_1 f_t - f_\infty\|^{2+\varepsilon} \leq \mathcal{E}(\Pi_1 f) - \mathcal{E}(f_\infty) \leq C_\varepsilon \|\Pi_1 f_t - f_\infty\|^{2-\varepsilon}$;

(iii) $\Pi_J f = f_\infty$; $Bf_\infty = 0$;

(H) $\|(\mathrm{Id} - \Pi_j)'_{\Pi_j f} \cdot (B \Pi_j f)\|^2 \geq K_\varepsilon \|(\Pi_j - \Pi_{j+1}) f\|^{2+\varepsilon}$ *for all* $j \in \{1, \dots, J-1\}$.

*Then* $\|f_t - f_\infty\| = O(t^{-\infty})$.

This theorem may seem particularly abstract and confusing, so I should give some explanations. First, $B$ plays the role of the antisymmetric part, but this shows only in the assumption that it does not contribute to the decay of $\mathcal{E}$; on the contrary, $\mathcal{C}$ should be thought of as the symmetric, or collisional part, and it does make the Lyapunov functional decay.

Next, the operators $\Pi_j$ act as a family of "nested projections". The first one, $\Pi_1$, sends $f$ to the kernel of the "collision operator" $\mathcal{C}$; then the second one sends $f$ to a smaller subspace, and then each $\Pi_j$ takes values in a smaller subspace until finally one reaches $f_\infty$. The "concrete" examples are the maps $f \to M^f_{\rho\, u\, T}$, $f \to M^f_{\rho\, u\, \langle T\rangle}$, $f \to M^f_{\rho\, 0\, 1}$, $f \to f_\infty$ which we considered in Section 2 (so for the Boltzmann equation we need four such nonlinear projections).

Finally, the key hypocoercivity condition is (H): It ensures basically that the effect of the "antisymmetric part" $B$ is strong enough to get us out of the image of $\Pi_j$, unless we are in the image of $\Pi_{j+1}$.

Theorem 3.9 leads to a simplified proof of Theorem 2.2, which does not involve any second-order differential inequality, but just variants of Gronwall's lemma. Once again, the key point is to add a correction to the Lyapunov functional $\mathcal{E}$ into another functional $\mathcal{F}$. The correction is small enough that the value of $\mathcal{F}$ is very close to the value of $\mathcal{E}$; but its structure is such that $\mathcal{F}$ satisfies (almost) a Gronwall-type estimate. More explicitly,

$$\mathcal{F}(f) = \big[\mathcal{E}(f) - \mathcal{E}(f_\infty)\big] + \sum_{j=1}^{J-1} a_j \big\langle (\mathrm{Id} - \Pi_j) f,\ (\mathrm{Id} - \Pi_j)'_f \cdot (Bf) \big\rangle, \tag{15}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in, say, $X^0$, and $\varepsilon > 0, a_j > 0\,(1 \le j \le J-1)$ are small numbers depending on the smoothness of $f_t$, on $\delta$, and on estimates on the distance of $f$ to $f_\infty$ (in general $1 \gg a_1 \gg \cdots \gg a_{J-1}$).

If the reader thinks that I am being too abstract and formal here, I invite him or her to write down explicitly what (15) is for the Boltzmann equation: Take $\mathcal{E}(f) = \int f \log f$, $f_\infty = M(v)$, $\Pi_1 f = M^f_{\rho\, u\, T}$, $\Pi_2 f = M^f_{\rho\, u\, \langle T \rangle}$, $\Pi_3 f = M^f_{\rho\, 0\, 1}$, $\Pi_4 f = f_\infty$, and $Bf = v \cdot \nabla_x f$; then the expression of $\mathcal{F}(f)$ would fill up basically a whole page. Expression (15) is not only quite general, it is also the best way to conduct calculations.

Let me conclude this section with another theorem that can be derived from Theorem 3.9: convergence to equilibrium for the nonlinear Vlasov–Fokker–Planck interaction with moderate interaction and small coupling.

**Theorem 3.10.** *Let $W \in C^\infty(\mathbb{T}^n)$ be an even smooth function with* $\sup W - \inf W$ *small enough, and let $f_0$ be a probability density on $\mathbb{T}^n_x \times \mathbb{R}^n_v$, with all moments finite. Then there is a unique solution $(f_t)_{t \ge 0}$ to the partial differential equation*

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f - \nabla_x (W * \rho) \cdot \nabla_v f = \Delta_v f + \nabla_v \cdot (f v), \quad \rho_t(x) = \int f_t(x, v)\, dv;$$

*and it does converge to a uniquely determined equilibrium distribution $f_\infty$, with*

$$\| f_t - f_\infty \|_{L^1} = O(t^{-\infty}).$$

This theorem also follows directly from Theorem 3.9, now by choosing just $\Pi_1 f = \rho M$, $\Pi_2 f = f_\infty$, where the equilibrium $f_\infty$ is the unique minimizer of the energy functional $H(f) + (1/2) \int \rho(x)\, \rho(y)\, W(x - y)\, dx\, dy$. The assumption on $W$ being smooth and small enough guarantees the uniqueness of the minimizer (it implies that we stay away from phase transitions) and allows to develop a very strong regularity theory for the equation.

## 4. Perspectives

There seems to be a whole mathematical world to explore behind the hypocoercivity phenomenon, both in nonlinear and linear partial differential equations.

Obvious remaining open problems concern the nonlinear equations such as Boltzmann's equation, for which the convergence to equilibrium is proven only under strong conditional smoothness assumptions; however, Theorem 2.2 says that, in some sense, it all amounts to a good understanding of the Cauchy problem. The situation is more subtle for coupled equations, such as the Vlasov–Fokker–Planck model: Theorem 3.10 solves the problem only for smooth *small enough* potentials, leaving completely open the issue of phase transition for large potentials. Realistic models such as the Vlasov–Poisson–Fokker–Planck equation require further thoughts.

But within the range of linear equations, where one is more demanding about conclusions, there is even much more to say. First, one would like to get a qualitative description of the convergence to equilibrium, and in particular of the oscillations described in the discussion of the Boltzmann equation. As discussed in [29], these oscillations appear in many models, but not always, and their presence or absence should be related to some spectral analysis of the linearized operator, involving conservation laws and hydrodynamical approximations. Recently, Francis Filbet, Clément Mouhot and Lorenzo Pareschi [10] have made some progress on this issue, by combining numerical simulations, linearization and asymptotic analysis; they suggest that for large domains the period of oscillations is given by the imaginary part of the eigenvalues of the linearized compressible Euler system, while the asymptotic rate of decay is determined by the real parts of the eigenvalues of the compressible linearized Navier–Stokes system. For small enough domains, the situation is completely different, and hydrodynamic effects should be negligible. Actually, numerical simulations in small domains show that the gas distribution first becomes spatially homogeneous, and then converges to equilibrium like a solution of the spatially homogeneous Boltzmann equation – a scenario which is somehow opposite to the ideas that seemed to be prevailing among physicists and mathematicians.

Another issue is about the quantitative relevance of the estimates. All the estimates derived from the hypocoercivity theorems in this text are explicit, but this does not mean that they have the correct order of magnitude. For simple equations like the Fokker–Planck equation, the rates of convergence predicted by my method seem to be off the true value by a factor of about $10^2$, which is not so bad (and much better than previous estimates). But the rates obtained for the oscillator chain have an incredibly bad dependence on the number of oscillators, leaving motivation for quantitative improvement.

It is important to note that the analysis of operators of the form $A^*A + B$ involved a systematic comparison with the symmetric operator $A^*A + [A, B]^*[A, B]$, or more complicated symmetric operators constructed from brackets with $B$. The same is true of the hypoelliptic method by Bernard Helffer and Francis Nier [16]. This might look satisfactory, but might also give wrong orders of magnitude. For instance, in the Model Problem 1.1, we have seen that if the eigenvalues for $A^*A + [A, B]^*[A, B]$ grow like $|\alpha|^{2\nu}$, then the real parts of the eigenvalues for $A^*A + B$ grow at least like $|\alpha|^{\nu}$. This behavior is optimal for some forms of the function $f$, but not for other ones, as pointed out to me by Thierry Gallay. Indeed, if $f(x) = 1/(1 + x^2)$, then $\nu = 1/4$, but numerical simulations suggest that the growth is like $|\alpha|^{1/2}\ldots$. This might indicate a fundamental limitation of present techniques, and motivate the development of a refined analysis.

In the example of the oscillator chain, the application of Theorem 3.4 is so far restricted to the oversimplified case when $T^{(\ell)} = T^{(r)}$. In fact, Theorem 3.4 could be applied if we had some basic qualitative information about the non-explicit equilibrium measure: For instance, some bounds on the Hessian of the logarithm of its density; and a Poincaré inequality. This leads to another challenging topic: deriving

qualitative *global* information about stationary solutions of linear partial differential equations.

Finally, the links and analogies between hypoellipticity and hypocoercivity need to be further explored. The interplay goes in both directions: It is possible to adapt some of the tricks presented here, into elementary methods for the study of hypoelliptic regularization. While these do not apply with such generality as the classical techniques introduced by Lars Hörmander and later by Joseph Kohn, they are quite flexible, in particular to get global estimates, or estimates from $L^1$ data. The same discovery has been made independently by Frédéric Hérau. Hopefully, all this agitation will lead to a new look at the old field of hypoelliptic regularity.

## References

[1] Cáceres, M. J., Carrillo, J. A., and Goudon, T., Equilibration rate for the linear inhomogeneous relaxation-time Boltzmann equation for charged particles. *Comm. Partial Differential Equations* **28** (5–6) (2003), 969–989.

[2] Carrillo, J. A., McCann, R. J., and Villani, C., Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoamericana* **19** (3) (2003), 971–1018.

[3] Desvillettes, L., and Villani, C., On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: the linear Fokker-Planck equation. *Comm. Pure Appl. Math.* **54** (1) (2001), 1–42.

[4] Desvillettes, L., and Villani, C., On a variant of Korn's inequality arising in statistical mechanics. *ESAIM Control Optim. Calc. Var.* **8** (2002), 603–619.

[5] Desvillettes, L., and Villani, C., On the trend to global equilibrium for spatially inhomogeneous kinetic systems: the Boltzmann equation. *Invent. Math.* **159** (2) (2005), 245–316.

[6] Eckmann, J.-P., and Hairer, M., Non-equilibrium statistical mechanics of strongly anharmonic chains of oscillators. *Comm. Math. Phys.* **212** (1) (2000), 105–164.

[7] Eckmann, J.-P., and Hairer, M., Spectral properties of hypoelliptic operators. *Comm. Math. Phys.* **235** (2) (2003), 233–253.

[8] Eckmann, J.-P., Pillet, C.-A., and Rey-Bellet, L., Non-equilibrium statistical mechanics of anharmonic chains coupled to two heat baths at different temperatures. *Comm. Math. Phys.* **201** (3) (1999), 657–697.

[9] Fellner, K., Neumann, L., and Schmeiser, C., Convergence to global equilibrium for spatially inhomogeneous kinetic models of non-micro-reversible processes. *Monatsh. Math.* **141** (4) (2004), 289–299.

[10] Filbet, F., Mouhot, C., and Pareschi, L., Solving the Boltzmann equation in $N \log_2 N$. *SIAM J. Sci. Comput.*, to appear.

[11] Gallay, T., and Wayne, C. E., Global stability of vortex solutions of the two-dimensional Navier-Stokes equation. *Comm. Math. Phys.* **255** (1) (2005), 97–129.

[12] Guo, Y., The Landau equation in a periodic box. *Comm. Math. Phys.* **231** (3) (2002), 391–434.

[13] Guo, Y., and Strain, R., Almost exponential decay near Maxwellian. *Comm. Partial Differential Equations* **31** (3) (2006), 417–429.

[14] Guo, Y., and Strain, R., Exponential decay for soft potentials near Maxwellian. Preprint, 2005.

[15] Hairer, M., and Mattingly, J., Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Ann. of Math*, to appear.

[16] Helffer, B., and Nier, F., *Hypoellipticity and spectral theory for Fokker-Planck operators and Witten Laplacians*, Lecture Notes in Math. 1862, Springer-Verlag, Berlin 2005.

[17] Hérau, F., Hypocoercivity and exponential time decay for the linear inhomogeneous relaxation Boltzmann equation. *Asymptot. Anal.*, to appear; http://helios.univ-reims.fr/Labos/Mathematiques/Homepages/Herau/.

[18] Hérau, F., Short and long time behavior of the Fokker-Planck equation in a confining potential and applications. Preprint (revised version), 2005; http://helios.univ-reims.fr/Labos/Mathematiques/Homepages/Herau/.

[19] Hérau, F., and Nier, F., Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Arch. Ration. Mech. Anal.* **171** (2) (2004), 151–218.

[20] Hörmander, L., Hypoelliptic second order differential equations. *Acta Math.* **119** (1967), 147–171.

[21] Mouhot, C., and Neumann, L., Quantitative perturbative study of convergence to equilibrium for collisional kinetic models in the torus. Preprint, 2006.

[22] Rey-Bellet, L., and Thomas, L. E., Asymptotic behavior of thermal nonequilibrium steady states for a driven chain of anharmonic oscillators. *Comm. Math. Phys.* **215** (1) (2000), 1–24.

[23] Rey-Bellet, L., and Thomas, L. E., Exponential convergence to non-equilibrium stationary states in classical statistical mechanics. *Comm. Math. Phys.* **225** (2) (2002), 305–329.

[24] Talay, D., Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Related Fields* **8** (2) (2002), 163–198.

[25] Toscani, G., and Villani, C., Sharp entropy dissipation bounds and explicit rate of trend to equilibrium for the spatially homogeneous Boltzmann equation. *Comm. Math. Phys.* **203** (3) (1999), 667–706.

[26] Toscani, G., and Villani, C., On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *J. Statist. Phys.* **98** (5–6) (2000), 1279–1309.

[27] Villani, C., A review of mathematical topics in collisional kinetic theory. In *Handbook of mathematical fluid dynamics*, Vol. I, North-Holland, Amsterdam 2002, 71–305.

[28] Villani, C., Cercignani's conjecture is sometimes true and always almost true. *Comm. Math. Phys.* **234** (3) (2003), 455–490.

[29] Villani, C., Entropy dissipation and convergence to equilibrium. Notes from a series of lectures at Institut Henri Poincaré, Paris 2001 (updated 2004); http://www.umpa.ens-lyon.fr/~cvillani/.

[30] Villani, C., Convergence to equilibrium: Entropy production and hypocoercivity. In *Rarefied Gas Dynamics* (ed. by M. Capitelli), AIP Conference Proceedings 762, American Institute of Physics, 2005, 8–25 .

[31] Villani, C., Hypocoercive diffusion operators in Hörmander form. Preprint, 2006; http://www.umpa.ens-lyon.fr/~cvillani/.

[32] Villani, C., Hypocoercive nonlinear diffusion operators. Preprint, 2006.

UMPA (UMR CNRS 5669), ENS Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France
E-mail: cvillani@umpa.ens-lyon.fr

# Metastability: a potential theoretic approach

Anton Bovier*

**Abstract.** Metastability is an ubiquitous phenomenon of the dynamical behaviour of complex systems. In this talk, I describe recent attempts towards a model-independent approach to metastability in the context of reversible Markov processes. I will present an outline of a general theory, based on careful use of potential theoretic ideas and indicate a number of concrete examples where this theory was used very successfully. I will also indicate some challenges for future work.

**Mathematics Subject Classification (2000).** Primary 60J45; Secondary 82C26.

**Keywords.** Metastability, Markov processes, potential theory, capacity, spectral theory.

## 1. Introduction

Metastability is a physical phenomenon that is observed in a large variety of situations in nature. The classical school-book example is the time delay in the evaporation of overheated water, resp. the delayed freezing of under-cooled water. Generally speaking, metastability is related to the existence of multiple, well separated time-scales: at a short time-scale, the systems *appears* to be in an equilibrium state, but really explores only a confined section of its available phase space, while, at much larger time scales, it undergoes transitions *between* such *metastable* states. The main mathematical task we want to discuss it the analysis of such system at these very long time-scales.

The mathematical description of metastable systems began in the 1930s and 1940s and is linked to the names of Eyring [15] and Kramer [26], who were interested in metastability in the context of chemical reactions. Kramer, in particular, introduced a one-dimensional diffusion process in a double-well potential as a model of a metastable system which is still used today in many applications. This work set the way to study metastability as a phenomenon that takes place in *stochastic processes*, and, in particular, in *Markov processes*, which is the setting that we will consider in this talk.

The mathematically rigorous analysis of metastability phenomena in the context of stochastic dynamics goes back essentially to the work of Freidlin and Wentzell in the early 1970s (see their seminal book [16]). They considered mainly the setting

*Support by the German Research Council (DFG) and by the European Science Foundation in the programme RDSES is gratefully acknowledged.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of finite dimensional dynamical systems perturbed by weak additive noise. In the simplest case, this would be driven by Brownian motion, but alternative settings, such as Levy-processes, were also considered. Metastability arises in this context if the unperturbed systems possesses several stable attractors. In this case, on short time-scales, the trajectories of the system will track those of the unperturbed system, and hence will converge towards one of the attractors. On much longer time-scales, however, the random perturbation allow the system to perform transitions between these stable attractors. The method to analyse the occurrence of such transitions introduced by Freidlin and Wentzell in this context was *large deviation theory* on path-space. This allows to control the probability of an "atypical" trajectory $\gamma(t)$ over some time interval $[T_1, T_2]$ in terms of an *action functional* $S(\gamma, T_1, T_2)$, which can be written as

$$S(\gamma, T_1, T_2) = \int_{T_1}^{T_2} \mathcal{L}(\gamma(t), \gamma'(t), t)\, dt \tag{1.1}$$

in the sense that

$$\lim_{\epsilon \downarrow 0} \epsilon \ln \mathbb{P}\left(X_\epsilon(t) \sim \gamma(t), \quad t \in [T_1, T_2]\right) = S(\gamma, T_1, T_2), \tag{1.2}$$

where $\epsilon$ is a parameter controlling the strength of the random perturbation, and $\mathcal{L}$ is a *Lagrangian* in the sense of classical mechanics. Clearly, such results allow to compute probabilities, and hence expected times, of the occurrence of transitions between attractors. We will refer to estimates of the type (1.2) as *logarithmic equivalence*. While these are in some sense rather crude estimates, the large deviation method has proven very useful due to its rather universal applicability. It has, in fact, dominated the field on the mathematical side, and large deviations and metastability are often seen as almost synonymous (see e.g. the recent monograph [35]).

Besides the large deviation estimates, Freidlin and Wentzell introduced a very interesting and useful way of looking at metastable systems by associating to it what they called a *Markov chain with exponentially small transition probabilities*. Here they associate to the original system a finite state Markov chain whose states label the different attractors of the underlying dynamical system. The transition probabilities of this new chain are then computed by finding the probability of the most likely trajectory linking two such attractors; by the foregoing discussion, this probability will be exponentially small. The long-time behaviour of the system, viewed on the coarse-grained level, will then be well-described by that of the associated finite Markov chain. It was later noted that such Markov chains arise also in other contexts, notably in stochastic dynamics of interacting particle systems at very low temperatures, and they have become a subject of intensive investigation in their own right ([33], [34], [10], [9]), initially again mainly through large deviation techniques.

In the physics literature, very early on more precise results than those provided by large deviation theory were sought. The modelling context here was mostly that of stochastic differential equations with small noise, i.e. the multi-dimensional extension of Kramer's approach. However, while in the one-dimensional case essentially

exact solutions are available, the multi-dimensional setting leads to partial differential equations that are not explicitly solvable, and those asymptotic analysis encountered considerable analytic difficulties. Several authors strove to overcome those and to derive asymptotic expansions in the parameter $\epsilon$ using methods similar to those used in the study of quantum mechanical tunneling (WKB-method); these results remained, however, on the formal level, as no error estimates could be proven. We will not enter the details of this development here, but refer to the excellent account given in [27]. Very recently, there has been renewed progress on this issue in the case of the reversible diffusions that we will comment on below [18], [19], [20].

Independent of the issue of rigour, the analytic approach has the disadvantage that it is applicable to a very limited class of models. Thus, more robust techniques that would still give precise results are sought for. Spectral theory for the generator of the Markov process appears as a natural tool, since long-term dynamical properties should find their encoding in the nature of the spectrum. In fact, the analysis of the *spectral gap* between the zero eigenvalue and the next-smallest eigenvalue of the generator has been a prominent topic in the theory of Markov processed, mainly as a tool to control convergence to equilibrium. Let us cite, from the vast body of literature, the papers [24], [31], [30]. The characterisation of metastability in terms of spectral properties was initiated in early work of Davies [11], [12], [13], and more recently continued by Gaveau and Schulman [17]. In view of numerical applications in dynamics of large bio-molecules, this issue was also addressed recently by Huisinga et al. [25].

Our own interest in the issue of metastability form the study of Gibbs distributions of disordered systems, and in particular the Hopfield model of neural networks. Here one is interested in the dynamical behaviour of Markov process on some high-dimensional state space, mostly chosen to be reversible with respect to a *Gibbs measure*. Considerable effort is invested in the analysis of the properties of these Gibbs measures in the limit of infinite dimensions. The question that then arises is what can be learned from these Gibbs measures, or, what do we need to know about the Gibbs measures in order to understand the long-term properties of the dynamics? This question can be seen as the leitmotiv behind our work. Other than that, we were aiming for an approach that would be to a large extend model independent, and that at the same time would provide finer estimates than those obtainable with large deviation methods. Our approach thus makes deliberate use of knowledge of the invariant measure, and, moreover, will always assume that the dynamics is reversible. This certainly leaves many interesting situations out of the reach of our methods, but still covers a range of important applications.

**Acknowledgements.** The work reported here is based on a joint effort with Véronique Gayrard, Michael Eckhoff, and Markus Klein to understand metastability. Applications of these ideas in concrete models have been worked out with Gerard Ben Arous, Alessandra Faggionato, Frank den Hollander, Francesco Manzo, and Francesca Nardi. I am deeply indebted to all these collaborators for their immense contributions.

## 2. Characterisations of metastability

The most general setting we will consider can be described as follows. We consider a Markov process $X_t$ on a measure space $\Gamma \supset X_t$ with discrete or continuous time $t$. We will usually assume that the process is uniquely ergodic with invariant measure $\mathbb{Q}$. We will denote the law of this process by $\mathbb{P}$. Moreover, we will denote by $\mathbb{P}_x$ the law of the process conditioned on $X_0 = x$. We will denote by $\tau_D$, $D \subset \Gamma$, the first entrance time of $X_t$ in $D$, i.e.

$$\tau_D \equiv \inf\{t > 0,\ X(t) \in D\}. \tag{2.1}$$

An intuitively appealing definition of metastability could be the following:

*A family of Markov processes is called metastable, if there exists a collection of disjoint sets $B_i \subset \Gamma$, such that*

$$\frac{\sup_{x \notin \cup_i B_i} \mathbb{E}_x \tau_{\cup_i B_i}}{\inf_{x \in \cup_i B_i} \mathbb{E}_x \tau_{\cup_{k \neq i} B_k}} = o(1). \tag{2.2}$$

Here $o(1)$ should be thought of as an intrinsic small parameter that characterises the "degree" of metastability. Often we will have to do with a family of processes indexed by a parameter, that allows to make (2.2) as small as we like.

Intuitively, this definition says that our process lingers around one of the subsets $B_i$ for a long time (resp. returns to $B_i$ many times) before it visits another of these sets, and so on. This can be re-expressed in a number of ways, e.g. in terms of the behaviour of empirical distributions, but we will not go into this.

This definition characterizes metastability in terms of physical properties, namely hitting times, of the system. The problem is that it is not immediately verifiable, since it involves mean hitting times, that are nor easy to compute. It would thus be desirable to have an equivalent definition involving more manageable quantities.

A further goal will be to derive further general properties of metastable systems. Since the definition implies frequent returns to the small starting set $B_i$ before transit to another set $B_j$, this suggests an exponential law for the transit times. This also suggests that we may expect to describe the process of successive visits to distinct $B_i$ asymptotically as a Markov process. The most fundamental result we want to achieve in this context is a characterization of the spectrum of the generator, resp. the transition matrix of a metastable process.

## 3. Markov processes and potential theory

Our approach to metastability relies heavily on some elementary potential theory for Markov processes. Let us briefly recall some basic facts and definitions. We will consider the case of discrete space and discrete time, but the same holds with obvious changes in the continuous setting.

Thus let $\Gamma$ be a discrete set, $\mathbb{Q}$ be a positive measure on $\Gamma$, and $P$ a stochastic matrix on $\Gamma$. We will denote by $-L$ the generator of the process in the case of continuous time, and set $L = \mathbb{1} - P$ in the case of discrete time[1]. We assume that $L$ is symmetric on the space $L^2(\Gamma, \mathbb{Q})$.

**Green's function.** Let $\Omega \subset \Gamma$. Consider for $\lambda \in \mathbb{C}$ and $g$ a real valued function on $\Omega$ the Dirichlet problem

$$\begin{aligned} (L-\lambda) f(x) &= g(x), \qquad x \in \Omega, \\ f(x) &= 0, \qquad x \in \Omega^c. \end{aligned} \tag{3.1}$$

Whenever $\lambda$ is such that the problem has a unique solution, then it can be expressed in terms of the Dirichlet Green's function $G^{\lambda}_{\Omega}(x, y)$ as $(L-\lambda)^{\Omega}$, i.e. for any $g \in C_0(\Omega)$,

$$f(x) = \sum_{y \in \Omega} G^{\lambda}_{\Omega}(x, y) g(y). \tag{3.2}$$

Recall that the spectrum of $L$ (more precisely the Dirichlet spectrum of the restriction of $L$ to $\Omega$, which we will sometimes denote by $L^{\Omega}$), is the complement of the set of values $\lambda$ for which $G^{\lambda}_{\Omega}$ defines a bounded operator.

**Equilibrium potential and equilibrium measure.** Let $A, D \subset \Gamma$. Then the $\lambda$-equilibrium potential $h^{\lambda}_{A,D}$ (of the capacitor $(A, D)$) is defined as the solution of the Dirichlet problem

$$\begin{aligned} (L-\lambda) h^{\lambda}_{A,D}(x) &= 0, \quad x \in (A \cup D)^c, \\ h^{\lambda}_{A,D}(x) &= 1, \quad x \in A, \\ h^{\lambda}_{A,D}(x) &= 0, \quad x \in D. \end{aligned} \tag{3.3}$$

Note that (3.3) has a unique solution provided $\lambda$ is not in the spectrum of $L^{(A \cup B)^c}$.

The equilibrium measure $e^{\lambda}_{A,D}$ is defined as the unique measure on $A$ such that

$$h^{\lambda}_{A,D}(x) = \sum_{y \in A} G^{\lambda}_{D^c}(x, y) e^{\lambda}_{A,D}(y). \tag{3.4}$$

(3.4) may also be written as

$$e^{\lambda}_{A,D}(y) = (L-\lambda) h^{\lambda}_{A,D}(y). \tag{3.5}$$

**Capacity.** We now restrict our attention to the case $\lambda = 0$. We write $h \equiv h^0$ and $e \equiv e^0$. The *capacity* of the capacitor $(A, D)$ is defined as

$$\operatorname{cap}(A, D) \equiv \sum_{y \in A} \mathbb{Q}(y) e_{A,D}(y). \tag{3.6}$$

[1] The choice of the sign is made so that $L$ is a positive definite operator.

Using (3.5) one derives after some algebra that

$$\mathrm{cap}(A, D) = \frac{1}{2}\sum_{x,y} \mathbb{Q}(y)p(x, y)\|h_{A,D}(x) - h_{A,D}(y)\|^2 \equiv \Phi(h_{A,D}), \tag{3.7}$$

where $p(x, y)$ are the transition probabilities (in discrete time) respectively transition rates (in discrete time). $\Phi$ is called the Dirichlet form (or energy) for the operator $L$.

A fundamental consequence of (3.7) is the variational representation of the capacity, namely

$$\mathrm{cap}(A, D) = \inf_{h\in\mathcal{H}_{A,D}} \Phi(h), \tag{3.8}$$

where $\mathcal{H}_{A,D}$ denotes the set of functions

$$\mathcal{H}_{A,D} \equiv \{h\colon \Gamma \to [0, 1] : h(x) = 0,\ x \in D,\ h(x) = 1,\ x \in A\}. \tag{3.9}$$

**Probabilistic interpretation.** If $\lambda = 0$, the equilibrium potential has a natural probabilistic interpretation in terms of hitting probabilities of this process, namely

$$h_{A,D}(x) = \mathbb{P}_x[\tau_A < \tau_D]. \tag{3.10}$$

The equilibrium measure has a nice interpretation in the discrete time case if $A = \{y\}$ is a single point:

$$e_{y,D}(y) = \mathbb{P}_y[\tau_D < \tau_y]. \tag{3.11}$$

If $\lambda \neq 0$, the equilibrium potential still has a probabilistic interpretation in terms of the Laplace transform of the hitting time $\tau_A$ of the process starting in $x$ and killed in $D$. Namely, we have for general $\lambda$ that

$$h^{\lambda}_{A,D}(x) = \mathbb{E}_x e^{\lambda\tau_A}\mathbb{1}_{\tau_A<\tau_D} \tag{3.12}$$

for $x \in (A \cup D)^c$, whenever the right-hand side is finite.

Note that (3.12) implies that

$$\frac{d}{d\lambda}h^{\lambda=0}_{A,D}(x) = \mathbb{E}_x\tau_A\mathbb{1}_{\tau_A<\tau_D}. \tag{3.13}$$

Differentiating the defining equation of $h^{\lambda}_{A,D}$ reveals that the function

$$w_{A,D}(x) = \begin{cases} \mathbb{E}_x\tau_A\mathbb{1}_{\tau_A<\tau_D}, & x \in (A \cup D)^c \\ 0, & x \in A \cup D \end{cases} \tag{3.14}$$

solves the inhomogeneous Dirichlet problem

$$Lw_{A,D}(x) = h_{A,D}(x), \quad x \in (A \cup D)^c, \tag{3.15}$$

$$w_{A,D}(x) = 0, \quad x \in A \cup D. \tag{3.16}$$

Therefore, the mean hitting time in $A$ of the process killed in $D$ can be represented in terms of the Green's function as

$$\mathbb{E}_x \tau_A \mathbb{1}_{\tau_A < \tau_D} = \sum_{y \in (A \cup D)^c} G_{(A\cup D)^c}(x, y) h_{A,D}(y). \tag{3.17}$$

Note that in the particular case when $D = \emptyset$, we get the familiar Dirichlet problem

$$L w_A(x) = 1, \quad x \in A^c, \tag{3.18}$$

$$w_A(x) = 0, \quad x \in A, \tag{3.19}$$

and the representation

$$\mathbb{E}_x \tau_A = \sum_{y \in A^c} G_{A^c}(x, y). \tag{3.20}$$

The full beauty of all this comes out when combining (3.4) with (3.17), resp. (3.20). Namely,

$$\begin{aligned} \mathbb{Q}(z)\mathbb{E}_z \tau_A e_{z,A}(z) &= \sum_{y \in A^c} \mathbb{Q}(y) G_{A^c}(y, z) e_{z,A}(z) \\ &= \sum_{y \in A^c} \mathbb{Q}(y) h_{z,A}(y) \end{aligned} \tag{3.21}$$

or

$$\mathbb{E}_z \tau_A = \frac{1}{\operatorname{cap}(z, A)} \sum_{y \in A^c} \mathbb{Q}(y) h_{z,A}(y) \tag{3.22}$$

**Remark 3.1.** Equation (3.22) relies explicitly on the discrete structure on the state space, or more precisely that for any $x \in \Gamma$, $\mathbb{Q}(x) > 0$. In the case of continuous state space, such formulas do not hold in the strict sense, or are not useful, but suitable "integral versions", involving integrals over suitably chosen small neighborhoods of e.g. the points $z$ in (3.22) are still valid, and can be used to more or less the same effect as the exact relations in the discrete case. This entails, however, some extra technical difficulties. In these notes we will therefore restrict our attention to the discrete case, where the principle ideas can be explained without being obscured by technicalities.

## 4. Capacitary characterization of metastability

The relation (3.22) between mean hitting times and capacities suggests an alternative characterisation of metastability through capacities. We will see that this entails many advantages.

**Definition 4.1.** Assume that $\Gamma$ is a discrete set. Then a Markov processes $X^t$ is $\rho$-metastable with respect to the set of points $\mathcal{M} \subset \Gamma$ if

$$\frac{\sup_{x \in \mathcal{M}} \operatorname{cap}(x, \mathcal{M}\backslash x)/\mathbb{Q}(x)}{\inf_{y \notin \mathcal{M}} \operatorname{cap}(y, \mathcal{M})/\mathbb{Q}(y)} \leq \rho \ll 1. \tag{4.1}$$

**Remark 4.2.** Definition 4.1 is useful since it involves quantities that are either "known", or expected to be easily controllable. It becomes intuitively more appealing if we notice that it can be written alternatively as

$$\frac{\sup_{x\in\mathcal{M}} \mathbb{P}_x[\tau_{\mathcal{M}\setminus x} < \tau_x]}{\inf_{y\notin\mathcal{M}} \mathbb{P}_y[\tau_{\mathcal{M}} < \tau_y]} \leq \rho \ll 1. \tag{4.2}$$

**Renewal estimates.** The estimation of the equilibrium through capacities is based on a renewal argument, that in the case of discrete state space is very simple.

**Lemma 4.3.** *Let $A, D \subset \Gamma$ be disjoint sets, and let $x \notin A \cup D$. Then*

$$h_{A,D}(x) \leq \min\left(\frac{\mathrm{cap}(x, A)}{\mathrm{cap}(x, D)}, 1\right). \tag{4.3}$$

**Remark 4.4.** Note that the power of Lemma 4.3 is more than doubled by judicious use of the elementary fact that $h_{A,D}(x) = 1 - h_{D,A}(x)$.

**Ultrametricity.** An important fact that allows to obtain general results under our definition of metastability is the fact that it implies approximate ultrametricity of capacities. This has been noted in [5].

**Lemma 4.5.** *Assume that $x, y \in \Gamma$ and $D \subset \Gamma$. If $\mathrm{cap}(y, D) \leq \delta \mathrm{cap}(y, x)$ for $0 < \delta < \frac{1}{2}$, then*

$$\frac{1-2\delta}{1-\delta} \leq \frac{\mathrm{cap}(x, D)}{\mathrm{cap}(y, D)} \leq \frac{1}{1-\delta}. \tag{4.4}$$

*Proof.* The proof of this lemma given in [5] is probabilistic and uses splitting and renewal ideas. It should be possible to prove this result with purely analytic arguments. □

Lemma 4.5 has the following immediate corollary, which is the version of the ultrametric triangle inequality we are looking for:

**Corollary 4.6.** *Let $x, y, z \in \mathcal{M}$. Then*

$$\mathrm{cap}(x, y) \geq \frac{1}{3} \min\left(\mathrm{cap}(x, z), \mathrm{cap}(y, z)\right) \tag{4.5}$$

In the sequel it will be useful to have the notion of a "valley" or "attractor" of a point in $\mathcal{M}$. We set for $x \in \mathcal{M}$,

$$A(x) \equiv \{z \in \Gamma \mid \mathbb{P}_z[\tau_x = \tau_{\mathcal{M}}] = \sup_{y\in\mathcal{M}} \mathbb{P}_z[\tau_x = \tau_{\mathcal{M}}]\} \tag{4.6}$$

Note that valleys may overlap, but from Lemma 4.5 it follows easily that the intersection has a vanishing invariant mass. The notion of a valley in the case of a diffusion process coincides with the intuitive notion.

**Mean times.** A very pleasant feature of the definition of metastability in terms of capacities is that it allows to relate some key capacities to mean hitting times in a very simple way.

**Theorem 4.7.** *Let $x \in \mathcal{M}$ and $J \subset \mathcal{M}\backslash x$ be such a that for all $m \notin J \cup x$ either $\mathbb{Q}(m) \ll \mathbb{Q}(x)$ or $\mathrm{cap}(m, J) \gg \mathrm{cap}(m, x)$. Then*

$$\mathbb{E}_x \tau_J = \frac{\mathbb{Q}(A(x))}{\mathrm{cap}(x, J)} (1 + o(1)) . \tag{4.7}$$

Finally we want to compute the mean time to reach $\mathcal{M}$ starting from a general point.

**Lemma 4.8.** *Let $z \notin \mathcal{M}$. Let $a \equiv \sup_{y \notin \mathcal{M}} \frac{\mathbb{Q}(y)}{\mathrm{cap}(y, \mathcal{M})}$. Then*

$$\mathbb{E}_z \tau_{\mathcal{M}} \leq a^{-2} |\Gamma|. \tag{4.8}$$

**Remark 4.9.** If $\Gamma$ is finite, the above estimate combined with Theorem 4.7 shows that the two definitions of metastability we have given in terms of mean times rep. capacities are equivalent. On the other hand, in the case of infinite state space $\Gamma$, we cannot expect the supremum over $\mathbb{E}_z \tau_{\mathcal{M}}$ to be finite, which shows that our first definition was somewhat naive. Note however that this case the estimate (4.8) can be improved by placing $|\Gamma|$ with $\sum_{y:\mathbb{Q}(y \leq \mathbb{Q}(z)} \mathbb{Q}(y)/\mathbb{Q}(z) + \sum_{y:\mathbb{Q}(y) > \mathbb{Q}(z)} 1$.

## 5. Spectral characterisation of metastability

We now turn to the characterisation of metastability through spectral data. We will show that Definition 4.1 implies that the spectrum of the generator decomposes into a cluster of $|\mathcal{M}|$ very small real eigenvalues that are separated by a gap from the rest of the spectrum.

**A priori estimates.** The first step of our analysis consists in showing that the matrix $L^{\mathcal{M}^c}$ (i.e. with Dirichlet conditions in all the points of $\mathcal{M}$) has a minimal eigenvalue that is not smaller than $O(a)$. This result needs sometimes some improvement, but is shows the basic twist. This is a simple application of a Donsker–Varadhan argument.

**Lemma 5.1.** *Let $\lambda^0$ denote the infimum of the spectrum of $L^{\mathcal{M}^c}$. Then*

$$\lambda^0 \geq \frac{1}{\sup_{x \in \Gamma} \mathbb{E}_z \tau_{\mathcal{M}}}. \tag{5.1}$$

**Remark 5.2.** Lemma 5.1 links the fast time scale to the smallest eigenvalue of the Dirichlet operator, as should be expected. Note that the relation is not very precise. We will soon derive a much more precise relation between times and eigenvalues for the cluster of small eigenvalues. As stated, it is useless in the case of infinite state

space. It can, however, be improved to give useful bounds under tightness conditions on $\mathbb{Q}$ [5].

**Characterization of small eigenvalues.** We will now obtain a representation formula for all eigenvalues that are smaller than $\lambda^0$. It is clear that there will be precisely $|\mathcal{M}|$ such eigenvalues. This representation was first exploited in [5], but already in 1973 Wentzell put forward very similar ideas.

The basic idea is to use the fact that the solution of the Dirichlet problem

$$\begin{aligned} (L-\lambda)f(x) &= 0, \qquad x \notin \mathcal{M}, \\ f(x) &= \phi_x, \qquad x \in \mathcal{M}, \end{aligned} \tag{5.2}$$

already solves the eigenvalue equation $L\phi(x) = \lambda\phi(x)$ everywhere except possibly on $\mathcal{M}$. The question if whether an appropriate choice of boundary conditions and the right choice of the value of $\lambda$ will actually lead to a solution. This is indeed the case.

**Lemma 5.3.** *Assume that $\lambda < \lambda^0$ is an eigenvalue of $L$ and $\phi(x)$ is the corresponding eigenfunction. Then the unique solution of* (5.2) *with $\phi_x = \phi(x)$, $x \in \mathcal{M}$, satisfies $f(y) = \phi(y)$, for all $y \in \Gamma$.*

Let us denote by $\mathcal{E}_{\mathcal{M}}(\lambda)$ the $|\mathcal{M}| \times |\mathcal{M}|$-matrix with elements

$$(\mathcal{E}_{\mathcal{M}}(\lambda))_{xy} \equiv e^{\lambda}_{z,\mathcal{M}\setminus z}(x). \tag{5.3}$$

**Lemma 5.4.** *A number $\lambda < \lambda^0$ is an eigenvalue of the matrix $L$ if and only if*

$$\det \mathcal{E}_{\mathcal{M}}(\lambda) = 0. \tag{5.4}$$

Anticipating that we are interested in small $\lambda$, we want to re-write the matrix $\mathcal{E}_{\mathcal{M}}$ in a more convenient form. To do so let us set

$$h^{\lambda}_x(y) \equiv h_x(y) + \psi^{\lambda}_x(y). \tag{5.5}$$

Then $\mathcal{E}(\lambda)$ can be written in the form

$$\begin{aligned} (\mathcal{E}_{\mathcal{M}}(\lambda))_{xz} = \mathbb{Q}(x)^{-1}\Bigg(&\frac{1}{2}\sum_{y\neq y'} \mathbb{Q}(y')p(y',y)[h_z(y') - h_z(y)][h_x(y') - h_x(y)] \\ &- \lambda \sum_y \mathbb{Q}(y)\left(h_z(y)h_x(y) + h_x(y)\psi^{\lambda}_z(y)\right)\Bigg) \end{aligned} \tag{5.6}$$

where the term involving $\psi^{\lambda}$ can be viewed as a more or less irrelevant

We are now in a position to relate the small eigenvalues of $L$ to the eigenvalues of the classical capacity matrix. Let us denote by $\|f\|_2$ the $\ell^2$-norm with respect to the measure $\mathbb{Q}$, i.e. $\|f\|_2^2 = \sum_y \mathbb{Q}(y)f(y)^2$.

**Theorem 5.5.** *If $\lambda < \lambda^0$ is an eigenvalue of $L$, then there exists an eigenvalue $\mu$ of the $|\mathcal{M}| \times |\mathcal{M}|$-matrix $\mathcal{K}$ whose matrix elements are given by*

$$\mathcal{K}_{zx} = \frac{\frac{1}{2}\sum_{y \neq y'} \mathbb{Q}(y')p(y', y)[h_z(y') - h_z(y)][h_x(y') - h_x(y)]}{\|h_z\|_2 \|h_x\|_2} \tag{5.7}$$

*such that $\lambda = \mu\,(1 + O(\rho(\epsilon))$.*

The computation of the eigenvalues of the capacity matrix is now in principle a finite, though in general not trivial problem. The main difficulty is of course the computation of the capacities and induction coefficients.

In fact we will prove the following theorem.

**Theorem 5.6.** *Assume that there exists $x \in \mathcal{M}$ such that for some $\delta \ll 1$*

$$\frac{\mathrm{cap}_x(\mathcal{M}\backslash x)}{\|h_x\|_2^2} \geq \delta \max_{z \in \mathcal{M}\backslash x} \frac{\mathrm{cap}_z(\mathcal{M}\backslash z)}{\|h_z\|_2^2}. \tag{5.8}$$

*Then the largest eigenvalue of $L$ is given by*

$$\lambda_x = \frac{\mathrm{cap}_x(\mathcal{M}\backslash x)}{\|h_x\|_2^2}(1 + O(\delta)) \tag{5.9}$$

*and all other eigenvalues of $L$ satisfy*

$$\lambda \leq C\delta\lambda_x. \tag{5.10}$$

*Moreover, the eigenvector $\phi$ corresponding to the largest eigenvalues normalized such that $\phi_x = 1$ satisfies $\phi_z \leq C\delta$, for $z \neq x$.*

Theorem 5.6 has the following simple corollary, that allows in many situations a complete characterization of the small eigenvalues of $L$.

**Corollary 5.7.** *Assume that we can construct a sequence of metastable sets $\mathcal{M}_k \supset \mathcal{M}_{k-1} \supset \cdots \supset \mathcal{M}_2 \supset \mathcal{M}_1 = x_0$, such that, for any $i$, $\mathcal{M}_i \backslash \mathcal{M}_{i-1} = x_i$ is a single point, and that each $\mathcal{M}_i$ satisfies the assumptions of Theorem* 5.6. *Then $L$ has $k$ eigenvalues*

$$\lambda_i = \frac{\mathrm{cap}_{x_i}(\mathcal{M}_{i-1})}{\mathbb{Q}(A(x_i))}\,(1 + O(\delta))\,. \tag{5.11}$$

*The corresponding normalized eigenfunction is given by*

$$\psi_i(y) = \frac{h_{x_i}(y)}{\|h_{x_i}\|_2} + O(\delta). \tag{5.12}$$

## 6. Variational principles and bounds for capacities

While the characterisation of metastability in terms of properties of mean first entrance times did not seem immediately verifiable in a given model, (3.22) already suggests a close relation between these times to capacities. In fact, the key idea in our approach will be to express all quantities of interest ultimately to capacities, and to exploit the fact that these can be estimated remarkably well by exploiting the variational principle (3.8). This observation is not new; in fact, it is the basis of the "electric network" approach to Markov chains (see e.g. the excellent account in Doyle and Snell [14]). The fact that this approach is very useful for metastable systems appears to have been overlooked.

Let us briefly comment on the use of these variational principles and explain why they are efficient. While the specifics of their exploitation are model-dependent, some basic principles are quite general follow directly from the fact that one is considering a metastable system.

**Upper bound.** Upper bounds on capacities can be gotten readily by judicious choices of a test-function $h$. Inspecting the Dirichlet form will often suggest rather good choices. There are two major advantages in this variational principle: there are no constraints on the test function except boundary conditions, and the minimizers has a very clear probabilistic interpretation. This is quite different from the situation of the Rayleigh–Ritz variational principle for the spectral gap, which is therefore more difficult to handle. The fact that a system is metastable suggests that there will be rather large regions, surrounding the metastable points, where the equilibrium potential is constant, and only its behaviour on the (often small) connecting sets has to be guessed with greater care.

**Lower bound.** A lower bound appears at first sight less obvious; however, the fact that $\Phi(h)$ is monotone in the variables $p(x, y)$ suggests an immediate lower bound in terms of the capacities of a chain there some (or even many) $p(x, y)$ are set to zero (known as Rayleigh's cut method [14]), hoping of course that the resulting chain will be so simple that explicit computations of the capacities are possible. This idea can, however, be extended considerably. To this end, consider a countable set $I$, and a let $\mathcal{G} \equiv \{g_{xy}, x, y \in \Gamma\}$, be a collection of sub-probability measures on $I$, i.e. for each $(x, y)$, $g_{xy}(\alpha) \geq 0$, and $\sum_{\alpha\in I} g_{xy}(\alpha) \leq 1$. Then

$$\begin{aligned}
\mathrm{cap}(A, B) &= \inf_{h\in\mathcal{H}_{A,D}} \sum_{\alpha\in I} \frac{1}{2} \sum_{x,y} \mathbb{Q}(y) g_{xy}(\alpha) p(x, y) \|h_{A,D}(x) - h_{A,D}(y)\|^2 \\
&\geq \sum_{\alpha\in I} \inf_{h\in\mathcal{H}_{A,D}} \sum_{\alpha\in I} \frac{1}{2} \sum_{x,y} \mathbb{Q}(y) g_{xy}(\alpha) p(x, y) \|h_{A,D}(x) - h_{A,D}(y)\|^2 \\
&\equiv \sum_{\alpha\in I} \inf_{h\in\mathcal{H}_{A,D}} \Phi^{\mathcal{G}(x)}(h) \equiv \sum_{\alpha\in I} \mathrm{cap}^{\mathcal{G}(\alpha)}(A, D).
\end{aligned} \tag{6.1}$$

As this it true for all $\mathcal{G}$, we get the variational principle

$$\operatorname{cap}(A, D) = \sup_{\mathcal{G}} \sum_{\alpha \in I} \operatorname{cap}^{\mathcal{G}(a)}(A, D). \tag{6.2}$$

Note that this may look trivial, as of course the supremum is realised for the trivial case $I = \{1\}$, $g_{xy}(1) = 1$, for all $(x, y)$. The interest in the principle arises from the fact that there may be other choices that still realise the supremum (or at least come very close to it). If we denote by $h_{A,D}^{\mathcal{G}(\alpha)}$ the minimizer of $\Phi^{\mathcal{G}(x)}(h)$, then $\mathcal{G}$ realises the supremum whenever

$$h_{A,D}^{\mathcal{G}(\alpha)}(x) = h_{A,D}(x), \quad \text{for all } x \text{ with } g(xy)(\alpha) \neq 0. \tag{6.3}$$

Of course we do not know $h_{A,D}(x)$, but this observation suggest a very good strategy to prove lower bounds, anyhow: guess a plausible test function $h$ for the upper bound, then try to construct $\mathcal{G}$ such that the minimizers, $h^{\mathcal{G}(\alpha)}$, are computable, and are similar to $h$! If this succeeds, the resulting upper and lower bounds will be at least very close. Remarkably, this strategy actually does work in many cases.

## 7. Applications 1. Low-temperature dynamics of spin systems

The somehow simplest example where the general approach outlined above works very well and with remarkable ease is stochastic dynamics of discrete spin systems in the low temperature limit.

Here we have a finite spin-space $\mathcal{S}$, a finite subset $\Lambda$ of $Z^d$, and a Hamiltonian function $H_\Lambda \colon \mathcal{S}^\Lambda \to \mathbb{R}$. The Markov processes one is interested in are reversible (discrete or continuous time) Markov chains on $\mathcal{S}^\Lambda$ that are reversible with respect to the *Gibbs measure*,

$$\mu_\beta(\sigma) = \frac{e^{-\beta H_\sigma(\sigma)}}{Z_{\beta,\Lambda}}, \tag{7.1}$$

where $\beta$ will play the rôle of a large parameter and $\mathbb{Z}_{\beta,\Lambda}$ is a normalisation constant, called partition function. To complete the description of the dynamics, one defines a graph $\Gamma = (\mathcal{S}^\Lambda, \mathcal{E}^\Lambda)$ on $\mathcal{S}^\Lambda$ whose edges determine the allowed transitions. One may then choose transition probabilities

$$p(\sigma, \sigma') = \frac{1}{C_\sigma} e^{-\beta[H_\Lambda(\sigma') - H_\Lambda(\sigma)]_+} \quad \text{if } (\sigma, \sigma') \in \mathcal{E}^\Lambda, \tag{7.2}$$

where $C_\sigma$ denotes the coordination number of the vertex $\sigma$ in $\Gamma$ and $[f]_+$ is the positive part of $f$; all other transitions have zero probability, except of course the probability to stay at $\sigma$, which is determined by the requirement that $p$ be a stochastic matrix. Such a dynamics is usually called a Metropolis algorithm.

Metastability occurs in such dynamics whenever the Hamiltonian has more than one local minimum, if $\beta$ is large. Our methods allow a full analysis of such dynamics,

provided we understand the function $H_\Lambda$ well enough to know its minima and saddle points. This latter problem is rather non-trivial in general and involves complicated discrete optimisation problems. In the Ising model, these have been studied in great detail by Alonso and Cerf [1]

The two most prominent examples in this class of models are the Glauber dynamics in the Ising spin model and Kawasaki dynamics in the Ising lattice gas.

**Ising model under Glauber dynamics.** Here the state space is $\mathcal{S} = \{-1, 1\}$, and the Hamiltonian is

$$H_\Lambda(\sigma) = -\sum_{x,y\in\Lambda,\|x-y\|=1} \sigma_x\sigma_y - h\sum_{x\in\Lambda}\sigma_x. \tag{7.3}$$

The edges of the graph $\Gamma$ of allowed transitions consists of all pairs $\sigma, \sigma'$) such that $\sigma$ and $\sigma'$ differ in exactly one coordinate, i.e. the Hamming distance between $\sigma$ and $\sigma'$ equals to 1. A detailed analysis of the structure of this Hamiltonian was given in [1], [9]. Here, if $h > 0$, the configurations $+\mathbf{1} \equiv \{\sigma_x \equiv +1 \text{ for all } x \in \Lambda\}$, and $-\mathbf{1} \equiv \{\sigma_x \equiv -1 \text{ for all } x \in \Lambda\}$, correspond to the deepest, resp. second-deepest minima. One can verify that the set $\mathcal{M} \equiv \{-\mathbf{1}, +\mathbf{1}\}$ is a set of metastable points in the sense of our definition. This dynamics was investigated in particular in [32] ($d = 2$) and [9] ($d = 2, 3$) using large deviation methods, and logarithmic asymptotics (in the regime $\Lambda$ fixed, $\beta \uparrow \infty$) were obtained, together with a description of the most probable exit path. In [4] we showed that the methods outlined above are readily applicable here and give a radical improvement on the precision of the results. We cite the main theorem of [4] to give a flavour of the type of results one can get.

**Theorem 7.1.** *Consider the Ising model with Metropolis dynamics in dimension $d = 2$ or $d = 3$ in a torus $\Lambda^d(l)$ with diameter $l$. Let $0 < h < 1$ such that $2/h$ and (in $d = 3$) $(h/2\lceil 4/h\rceil(4/h + 1 - \lceil 4/h\rceil)$ is not an integer. Then there exists $\delta > 0$, independent of $\beta$, such that the following hold.*

- *In dimension 2, let $h$ be such that $2/h$ is not an integer. Let $\ell_2 := \lceil \frac{2}{h} \rceil$ and $\Gamma_2 := 4\ell_2 - h(\ell_2^2 - \ell_2 + 1)$ be the diameter and the activation energy of the "critical droplet", respectively. Then*

$$\begin{aligned}\mathbb{E}\tau(-\mathbf{1}) &= \frac{3}{8}\frac{1}{\ell_2 - 1}e^{\beta\Gamma_2}(1 + O(e^{-\beta\delta}) \\ &= \frac{3}{16}he^{\beta\Gamma_2}(1 + O(h) + O(e^{-\beta\delta})).\end{aligned} \tag{7.4}$$

- *In dimension 3, let $h$ be such that $2/h$ and $(h/2\ \lceil 4/h\rceil\ (4/h + 1 - \lceil 4/h\rceil))$ are not integer. Let $\ell_3 := \lceil \frac{4}{h} \rceil$ and $a := \lceil h/2\ \lceil 4/h\rceil\ (4/h + 1 - \lceil 4/h\rceil)\rceil$ (notice that $a$ can take the value 1 or 2). The activation energy of the "critical droplet" is*

$$\begin{aligned}\Gamma_3 := (&6\ell_3^2 - (12 - 4a)\ell_3 + 4\ell_2 + 4 - 2a) \\ &- h(\ell_3^3 - (3 - a)\ell_3^2 + (2 - a) + \ell_2^2 - \ell_2 + 1).\end{aligned} \tag{7.5}$$

*Then*

$$\mathbb{E}\tau(-\mathbf{1}) = \frac{a}{16}\frac{1}{(\ell_3-\ell_2+1)(\ell_3-\ell_2+a-1)(\ell_2-1)}e^{\beta\Gamma_3}(1+O(e^{-\beta\delta})) = \frac{a}{128}h^3e^{\beta\Gamma_3}(1+O(h)+O(e^{-\beta\delta})). \tag{7.6}$$

*Moreover, the distribution of* $\tau(-\mathbf{1})/\mathbb{E}\tau(-\mathbf{1})$ *converges to the exponential distribution as* $\beta \uparrow \infty$.

**Local Kawasaki dynamics with open boundary conditions.** Kawasaki dynamics is most conveniently thought of a taking place on the configuration space $\{0,1\}^\Lambda$, where dynamics variable $\eta_x(t)$ is thought of as the number of particles at site $x$ at time $t$. In contrast to Glauber dynamics, Kawasaki dynamics is *conservative*, i.e. the total number of particles is fixed. In several papers, den Hollander et al. [21], [22] introduced a local version of this dynamics in a finite box where particle number of conserved by transitions within the box, but where particles may appear or disappear at the boundary. This dynamics was introduced as a local approximation of a true Kawasaki dynamics in infinite volume.

The Hamiltonian is written in these variables as

$$H_\Lambda(\eta) = -U\sum_{x,y\in\Lambda;\|x-y\|=1}\eta_x\eta_y - \Delta\sum_{x\in\Lambda}\eta_x. \tag{7.7}$$

Again the dynamics is chosen reversible with respect to the corresponding Gibbs measureand this time transition are possible between configurations $\eta, \eta'$ such that (i) either two nearest neighbor sites $x$, $y$ exchange their particle numbers, or (ii) the particle number of one site in the boundary of $\Lambda$ is increased from 0 to 1, or decreased from 1 to 0. This dynamics differs from Glauber dynamics essentially only in the structure of the graph of admissible transitions, but this has rather noticeable consequences.

From the point of view of metastability, the main new feature is that now the saddle points form Gauber dynamics have to be replaced by 'plateaus': there is a set of critical configurations where a "critical droplet" has been formed and a free particle has entered the box at the boundary. The droplet will become supercritical if this new particle attaches itself to the droplet. Therefore, in the computation of the capacities, the probability that a simple random walker starting at the boundary of a box $\Lambda$ will reach some set in the interior before exiting will play a crucial rôle and will in fact modify the prefactor of the nucleation time in a $\Lambda$ and dimension dependent way. Results analogous to those described in Section 7 have been obtained for Kawasaki dynamics in [8].

## 8. Applications 2. Diffusion processes

In the preceding sections we have explained our approach in the context of discrete Markov chains. A natural challenge was the extension of the methods to the classical examples of the theory of Freidlin and Wentzell, at least in the reversible case, and to consider stochastic differential equations of the type

$$dX_\epsilon(t) = -\nabla F(X_\epsilon(t))dt + \sqrt{2\epsilon}dW(t) \tag{8.1}$$

on a domain in $\mathbb{R}^d$, where $W$ is a $d$-dimensional Brownian motion, and $F$ a potential function (that shall satisfy suitable smoothness and boundary conditions). Note that this process is reversible with respect to the measure $\mathbb{Q}(dx) \equiv \exp(-F(x)/\epsilon)dx$.

The difficulties to overcome in this case is that, in dimension greater than 1, the process will not hit given points in finite time, and thus Definition 4.1 cannot be used. A viable alternative turns out to be:

**Definition 8.1.** Consider a diffusion process on a set $\Omega \subset \mathbb{R}^d$. The processes $X^t$ is $\rho$-metastable with respect to the set of points $\mathcal{M} \subset \Omega$ if

$$\frac{\sup_{x\in\mathcal{M}} \mathrm{cap}\,(B_\epsilon(x), \mathcal{M}\backslash x)\,/\mathbb{Q}\,(B_\epsilon(x))}{\inf_{z\in\Omega} \mathrm{cap}\,(B_\epsilon(z), \cup_{x\in\mathcal{M}}(B_\epsilon(x))\,/\mathbb{Q}\,((B_\epsilon(z)))} \leq \rho \ll 1, \tag{8.2}$$

where $B_\epsilon(x)$ denotes the Euclidean ball of radius $\epsilon$ around $x$.

Note that, if $F$ has finitely many local minima, one may chose the $\mathcal{M}$ as the set of local minima and $\rho = \exp(-c/\epsilon)$ for some $F$-dependent constant $a$.

This definition works, and in fact most of the approach of the discrete setting can be carried over to the diffusion case due to a-priori estimates on regularity properties of harmonic functions with respect to the generator of this process, which is the elliptic operator

$$-L_\epsilon \equiv -\epsilon\Delta - \nabla F(x)\cdot\nabla. \tag{8.3}$$

Elliptic regularity theory for local solutions of equation $L_\epsilon h = f$ implies that these tend to be Hölder continuous with constants those dependence on the small parameter can be controlled; typically they imply that the oscillation of such functions over balls of size $\epsilon$ is bounded by some positive power of $\epsilon$.

The estimates of capacities in the diffusion setup are again rather easily performed, and one obtains rigorous estimates of the prefactor of exit times from metastable sets as well as sharp asymptotics of the set of exponentially small eigenvalues associated with the local minima of $F$, under some non-degeneracy conditions (degenerate cases can in principle also be treated, but require special case by case analysis). The following result was proven in [6]:

**Theorem 8.2.** *Let $x_i$ be a minimum of $F$ and let $D$ be any closed subset of $\mathbb{R}^d$ such that:*

(i) *If $\mathcal{M}_i \equiv \{y_1, \ldots, y_k\} \subset \mathcal{M}$ enumerates all those minima of $F$ such that $F(y_j) \leq F(x_i)$, then $\cup_{j=1}^k B_\epsilon(y_j) \subset D$, and*

(ii) $\mathrm{dist}\,(\mathcal{S}(\mathrm{x_i}, \mathcal{M}_\mathrm{i}), \mathrm{D}) \geq \delta > 0$ *for some $\delta$ independent of $\epsilon$. Then*

$$\mathbb{E}_{x_i}\tau_D = \frac{2\pi\, e^{[F(z^*)-F(x_j)]/\epsilon}}{\sqrt{\det(\nabla^2 F(x_i))}\,\frac{|\lambda_1^*(z^*)|}{\sqrt{|\det(\nabla^2 F(z^*))|}}}\left(1 + O(\sqrt{\epsilon}|\ln\epsilon|)\right). \tag{8.4}$$

Here $z^*$ denotes the "minimal saddle point" between $x_i$ and the set $D$ (assumed unique), and $\lambda_i(z^*)$ denote the eigenvalues of the Hessian matrix of $\nabla^2 F(z^*)$ (assumed non-degenerate), $\lambda_1(z^*)$ being the unique negative one.

The following result on the small eigenvalues of $L_\epsilon$ is taken form [7]:

**Theorem 8.3.** *Assume that $F$ has $n$ local minima, $x_1, \ldots, x_n$ and that for some $\theta > 0$ the minima $x_i$ of $F$ can be labeled in such a way that, with $\mathcal{M}_k \equiv \{x_1, \ldots, x_k\}$ and $\mathcal{M}_0 \equiv \Omega^c$,*

$$F(z^*(x_k, \mathcal{M}_{k-1})) - F(x_k) \leq \min_{i<k}\left(F(z^*(x_i, \mathcal{M}_k \backslash x_i)) - F(x_i)\right) - \theta \tag{8.5}$$

*holds for all $k = 1, \ldots, n$. We will set $B_i \equiv B_\epsilon(x_i)$ and $\mathcal{S}_k \equiv \bigcup_{i=1}^k B_i$, and $h_k(y) \equiv h_{B_k, \mathcal{S}_{k-1}}(y)$. Assume moreover that all saddle points $z^*(x_k, \mathcal{M}_{k-1})$ are unique, and that the Hessian of $F$ is non-degenerate at all these saddle points and at all local minima. Then there exists $\delta > 0$ such that the $n$ exponentially small eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ of $L_\epsilon$ satisfy*

$$\lambda_1 = 0, \tag{8.6}$$

*and for $k = 2, \ldots, n$,*

$$\begin{aligned}
\lambda_k &= \frac{\mathrm{cap}_{B_k}(\mathcal{S}_{k-1})}{\|h_k\|_2^2}(1 + O(e^{-\delta/\epsilon})) \\
&= \frac{1}{\mathbb{E}_{x_k}\tau_{\mathcal{S}_{k-1}}}(1 + O(e^{-\delta/\epsilon})) \\
&= \frac{|\lambda_1^*(z^*(x_k, \mathcal{M}_{k-1}))|}{2\pi}\sqrt{\frac{\det(\nabla^2 F(x_k))}{|\det(\nabla^2 F(z^*(x_k, \mathcal{M}_{k-1})))|}}\, e^{-[F(z^*(x_k, \mathcal{M}_{k-1}))-F(x_k)]/\epsilon} \\
&\quad \times \left(1 + O(\epsilon^{1/2}|\ln\epsilon|)\right)
\end{aligned} \tag{8.7}$$

*where $\lambda_1^*(z^*)$ denotes the unique negative eigenvalue of the Hessian of $F$ at the saddle point $z^*$.*

**Remark 8.4.** After the results of [7] appeared, Helffer and Nier [19] reconsidered earlier work of Helffer and Sjösstrand concerning spectral asymptotics of Schrödinger operators with multi-well potentials. In two papers [18], [20] they showed that using the so-called Witten complex, it is possible to derive similar spectral results and even extend them to complete asymptotic expansions, provided $F$ is assumed smooth. This is an interesting alternative approach, that for the time being has the disadvantage to be limited to the diffusion setting.

## 9. Challenges

The main challenges to the approach to metastability outlined in this talk are models in very high, respectively infinite, dimensions. The most interesting examples here are stochastic dynamics of spin systems beyond the low-temperature regime and in large or infinite volume. There are two major difficulties that present themselves. The first one is the estimation of capacities, and more precisely the lower bounds. The second problem is that of proving some a priori regularity properties, similar to the case of diffusion processes. Both issues seem at the moment quite open and will probably require the analysis of model problems. If these problems can be understood, we would feel that metastability is rather well understood in the context of reversible Markov processes. An altogether different and wide open issue is metastability in non-reversible systems. Here, the theory of Freidlin and Wentzell remains the only generally applicable tool, and finding ways to get estimates of higher precision than those obtainable from large deviation theory remains, at least in general, unsolved.

## References

[1] Alonso, L., and Cerf, R., The three-dimensional polyominoes of minimal area. *Electron. J. Combin.* **3** (1996), Research Paper 27 (electronic).

[2] Bovier, A., Metastability and ageing in stochastic dynamics. In *Dynamics and randomness II* (A. Maas, S. Martínez, J. San Martin, eds.), Nonlinear Phenomena and Complex Systems 10, Kluwer Academic Publishers, Dordrecht 2004, 17–80.

[3] Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M., Metastability in stochastic dynamics of disordered mean-field models. *Probab. Theory Related Fields* **119** (2001), 99–161.

[4] Bovier, A., Manzo, F., Metastability in Glauber dynamics in the low-temperature limit: beyond exponential asymptotics. *J. Statist. Phys.* **107** (2002), 757–779.

[5] Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M., Metastability and low lying spectra in reversible Markov chains. *Comm. Math. Phys.* **228** (2002), 219–255.

[6] Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M., Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc.* (*JEMS*) **6** (2004), 399–424.

[7] Bovier, A., Gayrard, V., and Klein, M., Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc.* (*JEMS*) **7** (2005), 69–99.

[8] Bovier, A., den Hollander, F., and Nardi, F. R., Sharp asymptotics for Kawasaki dynamics on a finite box with open boundary conditions. *Probab. Theor. Rel. Fields.* **135** (2) (2006), 265–310 .

[9] Ben Arous, G., and Cerf, R., Metastability of the three-dimensional Ising model on a torus at very low temperatures. *Electron. J. Probab.* **1** (1996), 55 pp. (electronic).

[10] Catoni, O., and Cerf, R., The exit path of a Markov chain with rare transitions. *ESAIM Probab. Statist.* **1** (1995/97), 95–144 (electronic).

[11] Davies, E. B., Metastable states of symmetric Markov semigroups. I. *Proc. Lond. Math. Soc.* (3) **45** (1982), 133–150 .

[12] Davies, E. B., Metastable states of symmetric Markov semigroups. II. *J. Lond. Math. Soc.* (2) **26** (1982), 541–556.

[13] Davies, E. B., Spectral properties of metastable Markov semigroups. *J. Funct. Anal.* **52** (1983), 315–329.

[14] Doyle, P. G., and Snell, J. L., *Random walks and electrical networks*. Carus Mathematical Monographs 22, Mathematical Association of America, Washington, DC, 1984.

[15] Eyring, H., The activated complex in chemical reactions. *J. Chem. Phys.* **3** (1935), 107–115.

[16] Freidlin, M. I., and Wentzell, A. D., *Random perturbations of dynamical systems.* Second edition, Grundlehren Math. Wiss. 260, Springer-Verlag, New York 1998.

[17] Gaveau, B., and Schulman, L. S., Theory of nonequilibrium first-order phase transitions for stochastic dynamics. *J. Math. Phys.* **39** (1998), 1517–1533.

[18] Helffer, B., Klein, M., and Nier, F., Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Mat. Contemp.* **26** (2004), 41–85.

[19] Helffer, B., and Nier, F., *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, Lecture Notes in Math. 1862, Springer-Verlag, Berlin 2005.

[20] Helffer, B., and Nier, F., Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. The case with boundary. Preprint 04-40 IRMAR, Université de Rennes I, 2004.

[21] den Hollander, F., Olivieri, E., and Scoppola, E., Metastability and nucleation for conservative dynamics. Probabilistic techniques in equilibrium and non-equilibrium statistical physics. *J. Math. Phys.* **41** (2000), 1424–1498.

[22] den Hollander, F., Nardi, F. R., Olivieri, E., and Scoppola, E., Droplet growth for three-dimensional Kawasaki dynamics. *Probab. Theory Related Fields* **125** (2003), 153–194.

[23] den Hollander, F., Metastability under stochastic dynamics. *Stochastic Process. Appl.* **114** (2004), 1–26.

[24] Holley, R. A., Kusuoka, S., and Stroock, W.S., Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.* **83** (1989), 333–347.

[25] Huisinga, W., Meyn, S., and Schütte, Ch., Phase transitions and metastability for Markovian and molecular systems. *Ann. Appl. Probab.* **14** (2004), 419–458.

[26] Kramers, H. A., Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7** (1949), 284–304.

[27] Maier, R. S. and Stein, D. L., Limiting exit location distributions in the stochastic exit problem. *SIAM J. Appl. Math.* **57** (1997), 752–79.

[28] Martinelli, F., On the kinetic Ising model below the critical temperature. In *XIIIth International Congress on Mathematical Physics* (London, 2000), International Press, Boston, MA, 2001, 297–301.

[29] Martinelli, F., Relaxation times of Markov chains in statistical mechanics and combinatorial structures. In *Probability on discrete structures*, Encyclopaedia Math. Sci. 110, Springer-Verlag, Berlin 2004, 175–262.

[30] Mathieu, P., Spectra, exit times, and long times asymptotics in the zero white noise limit. *Stoch. Stoch. Rep.* **55** (1995), 1–20.

[31] Miclo, L., Comportement de spectres d'opérateurs de Schrödinger à basse température. *Bull. Sci. Math.* **119** (1995), 529–553.

[32] Neves, E. J., and Schonmann, R. H., Critical droplets and metastability for a Glauber dynamics at very low temperature. *Commun. Math. Phys.* **137** (1991), 209–230.

[33] Olivieri, E., and Scoppola, E., Markov chains with exponentially small transition probabilities: first exit problem from a general domain. I. The reversible case. *J. Statist. Phys.* **79** (1995), 613–647.

[34] Olivieri, E., and Scoppola, E., Markov chains with exponentially small transition probabilities: first exit problem from a general domain. II. The general case. *J. Statist. Phys.* **84** (1996), 987–1041.

[35] Olivieri, E., and Vares, M. E., *Large deviations and metastability.* Encyclopedia of Mathematics and its Applications 100, Cambridge University Press, Cambridge 2005.

Weierstrass-Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, and
Mathematics Institute, Technical University Berlin, Strasse des 17. Juni 136, 10623 Berlin, Germany

E-mail: bovier@wias-berlin.de

# On Ising droplets

Raphaël Cerf

**Abstract.** One of the fundamental goals of statistical mechanics is to understand the macroscopic effects induced by the random forces acting at the microscopic level. Some satisfactory results are now available for the Ising model at equilibrium in the phase coexistence regime in any dimension: it is rigorously proved that the most likely shapes of the macroscopic droplets of one pure phase floating in the other pure phase are close to the Wulff crystal of the model. However, the dynamical processes leading to the emergence of a droplet are far from being understood. We formulate a classical conjecture: the scaling limit of the Glauber microscopic dynamics should be an anisotropic motion by mean curvature.

**Mathematics Subject Classification (2000).** Primary 82B20; Secondary 82C20.

**Keywords.** Ising model, Wulff crystal, Glauber dynamics, mean curvature motion.

## 1. Water and oil

Let us consider a volume of water in absence of gravity at ordinary temperature. We start to pour a very small quantity of oil into the water. First, nothing noticeable happens on the macroscopic scale, i.e., the oil is perfectly dissolved throughout the water and the oil molecules are homogeneously spread within the water: by observing the liquid at the macroscopic level, we cannot even tell that it is a mixture of two distinct types of particles, which nevertheless have the tendency to repel each other. Let us keep pouring oil into the water. We know that the solubility of oil in water is not infinite; at some density threshold (which increases with the temperature), we obtain a solution of water saturated with oil. This solution is still a pure phase, completely homogeneous on the macroscopic level, and it realizes a perfect tradeoff between entropy and energy; we call it the water phase. Let us pour in some more oil. The excess of oil is not dissolved any more and it precipitates: macroscopic droplets of oil emerge. These droplets are not regions where there are only oil molecules, rather in these regions we observe the symmetric pure phase consisting of oil saturated with water, which we call the oil phase. The droplets are delimited by an abrupt change of the local density of water and oil molecules. We wish to understand the law governing the evolution and the shapes of these droplets.

The classical phenomenological theory asserts the existence of a macroscopic surface free energy $\mathcal{I}$ and that the droplets evolve so as to minimize $\mathcal{I}$. For instance, at equilibrium, in case $\mathcal{I}$ is isotropic, one observes a unique spherical droplet of the

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

oil phase floating in the sea of the water phase. Our aim is to confirm the predictions of the phenomenological theory starting from a truly microscopic model. We wish to understand how the random forces acting at the atomic level, more precisely the probabilistic repulsive effect between the two types of particles, can induce such deterministic macroscopic effects. One of the most famous results in probability theory is the law of large numbers: If $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent identically distributed random variables with mean $m$, then, with probability one,

$$\lim_{n\to\infty} \frac{1}{n}(X_1 + \cdots + X_n) = m.$$

What we have in mind is a generalization of the law of large numbers, but in a fundamentally new context, which we could state informally as follows:

$$\lim_{\text{number of particles}\to\infty} \begin{pmatrix} \text{global effect of the random} \\ \text{microscopic repulsive forces} \end{pmatrix} = \text{single droplet.}$$

The limiting deterministic object is the shape of the droplet at equilibrium and the problem is intrinsically geometric; we deal with spatially dependent random variables and we leave radically and definitively the independent framework. Hence the geometry enters the problem in a decisive way, in the random interactions and in the formulation of the result itself.

Let us try to set up a simple model of our experiment with water and oil. A convenient choice is a lattice model: each site of the lattice is occupied either by a water particle or by an oil particle, which we indicate respectively by $+$ or $-$. The interaction between different particles is repulsive and occurs when the substances are in immediate contact. Hence a repulsive nearest neighbour interaction is a sensible choice. Since we focus only on the repulsive interaction between different molecules, we can assume that the two substances are symmetric and that their self-interactions are of equal magnitude, or equivalently, equal to zero. We do not assume that the self-interactions between two particles of the same type are negligible compared to the repulsive effect; rather, we say that because of the symmetry, the global effect of the self-interactions cancels out. Thus the total energy of a configuration should be simply the number of all nearest neighbour pairs with different signs. We end up exactly with the Hamiltonian of the famous Ising model (to be defined precisely in the next section). In our experiment the density of oil is fixed, therefore we have a constraint on the possible configurations: the proportion of pluses and minuses has to be fixed. This situation amounts to considering the Ising model with plus boundary conditions (guaranteeing the water dominance) conditioned on the event that the average magnetization is equal to a fixed value smaller than the spontaneous magnetization at the given temperature.

## 2. Definition of the Ising model

For reasons of technical simplicity, it is easier to build our model on a lattice. We will work with the lattice $\mathbb{Z}^d$; each site of the lattice is occupied by one of the two types of particles that we denote by $-$ and $+$. Let $\Lambda \subset \mathbb{Z}^d$ be a cubic box. A configuration in $\Lambda$ is a map $\sigma : \Lambda \to \{-, +\}$ and for $x \in \Lambda$, we denote by $\sigma(x)$ the type of the particle present at $x$. The energy or Hamiltonian $H_\Lambda(\sigma)$ of the configuration $\sigma$ in $\Lambda$ is, up to a constant, twice the number of interfaces between the minuses and the pluses, that is,

$$H_\Lambda(\sigma) = -\frac{1}{2} \sum_{\substack{x,y\in\Lambda \\ |x-y|=1}} \sigma(x)\sigma(y) = - \sum_{\substack{\{x,y\}\in\Lambda^2 \\ |x-y|=1}} \sigma(x)\sigma(y).$$

We use the standard rules to multiply signs: $++ = -- = +$, $-+ = +- = -$. The first sum is above ordered pairs (whence the factor $1/2$) while the second is above unordered pairs. We need also a mechanism to ensure the dominance of one type of particles. This is achieved through boundary conditions. We consider only two types of boundary conditions, by putting either a layer of pluses or of minuses around the box $\Lambda$. The energy or Hamiltonian $H^*_\Lambda(\sigma)$ with boundary conditions $*$ (where $*$ stands for $-$ or $+$) is defined as above for the configurations $\sigma$ such that $\sigma(x) = *$ for all the sites $x$ in $\Lambda$ which are at a distance less than or equal to 1 from the complement of $\Lambda$ and $H^*_\Lambda(\sigma) = +\infty$ otherwise. Next we add some randomness in the model. Let $T > 0$ be the temperature. We build a probability law on the space $\{-, +\}^\Lambda$ of the configurations. This space is huge but finite, hence to define the law we need to specify the individual probability of each possible configuration. The natural way to do this is to use the Boltzmann factor. So, the Gibbs measure $\mu^*_{\Lambda,T}$ in $\Lambda$ at temperature $T$ with boundary conditions $*$ is given by

$$\mu^*_{\Lambda,T}(\sigma) = \frac{1}{Z^*_{\Lambda,T}} \exp -\frac{H^*_\Lambda(\sigma)}{T} \quad \text{for all } \sigma \in \{-, +\}^\Lambda,$$

where the normalizing factor $Z^*_{\Lambda,T}$, called the partition function, is equal to

$$Z^*_{\Lambda,T} = \sum_{\sigma\in\{-,+\}^\Lambda} \exp -\frac{H^*_\Lambda(\sigma)}{T}.$$

Whenever the superscript $*$ is absent, the boundary conditions are not specified and we have the Ising Gibbs measure $\mu_{\Lambda,T}$ with free boundary conditions, associated to the Hamiltonian $H_\Lambda$. Let us take a closer look at this formula. The elements $\Lambda$, $T$, $* \in \{-, +\}$ being fixed, the most likely configurations are those having a small energy, i.e., those for which the contacts between the minuses and the pluses are reduced. Thus we have built a complex probability law with strong spatial correlations. We shall next play a bit with the elements controlling the Gibbs measures $\mu^\pm_{\Lambda,T}$ in order to get some feeling for their influence.

**First asymptotics.** Imagine that we fix the box $\Lambda$ and that we set the boundary conditions to $+$. If we send $T$ to 0, then the measure $\mu^+_{\Lambda,T}$ concentrates on the configuration which realizes the global minimum of the Hamiltonian $H^+_\Lambda$, in this case the configuration where all the sites are pluses. On the contrary, if we send $T$ to $\infty$, the value of the Hamiltonian becomes irrelevant and $\mu^+_{\Lambda,T}$ converges towards the Bernoulli product law where all the sites are independent. The case of $\mu^-_{\Lambda,T}$ being symmetric, we see that

$$\begin{array}{rcccl} \text{Dirac mass at ``all pluses''} & \longleftarrow & \mu^+_{\Lambda,T} & \longrightarrow & \text{i.i.d. Bernoulli} \\ & T \downarrow 0 & & T \uparrow \infty & \\ \text{Dirac mass at ``all minuses''} & \longleftarrow & \mu^-_{\Lambda,T} & \longrightarrow & \text{i.i.d. Bernoulli} \end{array}$$

Something remarkable has already happened: as $T \uparrow \infty$, the boundary conditions are forgotten, while as $T \downarrow 0$, they completely determine the limit. However, we wish to work at a fixed positive temperature $T$. In order to observe a sharp mathematical phenomenon, we consider another kind of limit, namely the thermodynamic limit where the number of particles goes to infinity. This is achieved by letting the box $\Lambda$ grow and invade the whole lattice $\mathbb{Z}^d$. As $\Lambda$ increases to $\mathbb{Z}^d$, the expectation $\mu^+_{\Lambda,T}(\sigma(0))$ decreases and converges towards a limiting quantity $m^*(T)$:

$$\lim_{\Lambda \uparrow \mathbb{Z}^d} \mu^+_{\Lambda,T}(\sigma(0)) = m^*(T) = -\lim_{\Lambda \uparrow \mathbb{Z}^d} \mu^-_{\Lambda,T}(\sigma(0)).$$

Here is a heuristic explanation for this monotone convergence. Let us consider a huge box $\Lambda$ and the site at the center of the box $\Lambda$. With free boundary conditions, the law of $\sigma(0)$ under $\mu_{\Lambda,T}$ is symmetric, hence it is the one of a fair coin, i.e.,

$$\mu_{\Lambda,T}(\sigma(0) = +) = 1/2 = \mu_{\Lambda,T}(\sigma(0) = -).$$

If we put $+$ boundary conditions, these boundary conditions start to influence positively the sites at distance 1 from the boundary of the box, which themselves influence the sites at distance 2 from the boundary. This effect propagates and reaches the origin, so that the law of $\sigma(0)$ under $\mu^+_{\Lambda,T}$ is slightly biased towards $+$:

$$\mu^+_{\Lambda,T}(\sigma(0) = +) \; > \; 1/2 \; > \; \mu^+_{\Lambda,T}(\sigma(0) = -).$$

The larger the box $\Lambda$ is, the smaller is the resulting effect at the origin, hence the influence of the boundary conditions decreases as the box increases and the following monotone limit exists:

$$m^*(T) = \lim_{\Lambda \uparrow \mathbb{Z}^d} \mu^+_{\Lambda,T}(\sigma(0)).$$

The fundamental and basic question is whether something of the influence of the boundary conditions still remains after we have sent them to infinity. Equivalently, is $m^*(T)$ equal to 0?

The quantity $m^*(T)$ is called the spontaneous magnetization at temperature $T$. This terminology stems from the fact that the Ising model was originally introduced

as a model of ferromagnetism: under some adequate conditions, a magnet submitted to the influence of a magnetic field will remember the sign of the field even after it has disappeared (see [24] and the references therein for a serious physical introduction to the Ising model).

**Phase transition.** We say that there is a phase transition at temperature $T$ if $m^*(T) > 0$. The first fundamental result concerning the phase transition in the Ising model is the following.

**Theorem 2.1.** *In any dimension $d \geq 2$, there exists a positive and finite critical temperature $T_c(d)$ such that the Ising model exhibits a phase transition for $T < T_c(d)$ and it does not for $T > T_c(d)$.*

It is also possible to take the thermodynamic limit of the finite volume Gibbs measure $\mu^+_{\Lambda,T}$, and not only of the expected value $\mu^+_{\Lambda,T}(\sigma(0))$. As $\Lambda$ increases to $\mathbb{Z}^d$, the measure $\mu^+_{\Lambda,T}$ decreases stochastically and converges weakly towards the infinite volume Gibbs measure $\mu^+_T$, which is a probability measure on the space of infinite volume configurations $\{-,+\}^{\mathbb{Z}^d}$. Similarly, $\mu^-_{\Lambda,T}$ increases weakly towards a measure $\mu^-_T$:

$$\lim_{\Lambda\uparrow\mathbb{Z}^d} \mu^-_{\Lambda,T} = \mu^-_T, \quad \lim_{\Lambda\uparrow\mathbb{Z}^d} \mu^+_{\Lambda,T} = \mu^+_T.$$

The spontaneous magnetization $m^*(T)$ is equal to the expected value of $\sigma(0)$ under $\mu^+_T$ and there is a phase transition at temperature $T$ if and only if $\mu^-_T$ and $\mu^+_T$ are distinct. In other words, we have

$$m^*(T) > 0, \quad \mu^-_T \neq \mu^+_T \qquad \text{for all } T < T_c(d),$$

whereas for $T > T_c(d)$, we have $m^*(T) = 0$ and $\mu^-_T = \mu^+_T$.

## 3. The Wulff crystal

We shall mimic mathematically the initial experiment of Section 1 with the help of the Ising model. Let us consider a box $\Lambda(n)$ of diameter $n$ full of pluses. We take $n$ very large, of the order of the Avogadro number $6.02 \times 10^{23}$. We start deleting pluses and replacing them by minuses, first a small quantity of minuses. It is possible to build a stochastic dynamics in the box which is conservative (i.e., the total numbers of minuses and pluses remain unchanged or equivalently the empirical magnetization $n^{-d}\sum_{x\in\Lambda(n)}\sigma(x)$ remains constant) and whose final equilibrium is the Gibbs measure $\mu^+_{\Lambda(n),T}$ conditioned to have the initial fixed magnetization. The simplest such dynamics is the so-called Kawasaki dynamics: at random exponential times, a pair of neighbouring particles might be exchanged according to a simple local probabilistic rule (see Section 4). As long as the empirical magnetization is larger than $m^*(T)$, the configuration in $\Lambda(n)$ at equilibrium is expected to be spatially homogeneous. If

we keep pouring minuses into the box and removing pluses, we soon reach the value $m^*(T)$, and at this point we obtain the saturated pure phase $\mu_T^+$, i.e., the configuration in $\Lambda(n)$ looks like a finite sample of the infinite volume Gibbs measure $\mu_T^+$. We finally add some more minuses and we cross the threshold $m^*(T)$. We wish to understand the response of the system and the most likely configurations inside the box when there is an excess of minuses.

It turns out that this simple model indeed confirms the prediction of the phenomenological theory. At equilibrium, with probability tending to 1 as $n$ goes to $\infty$, a region emerges inside the box $\Lambda(n)$ where the configuration statistically looks like the minus phase $\mu_T^-$, surrounded by a region filled with the plus phase $\mu_T^+$. When rescaled by a factor $n$, the shape of this region converges as $n$ goes to $\infty$ towards a deterministic shape, called the Wulff crystal of the Ising model. This crystal is convex, it depends on the temperature and on the initial lattice $\mathbb{Z}^d$; it bears the name of Wulff, who studied it one century ago [33].

In order to detect conveniently the Wulff region, we rescale the box $\Lambda(n)$ by a factor $n$ and we send it onto the $d$-dimensional unit cube $[-1/2, 1/2]^d$. Let $\sigma$ be a spin configuration in $\Lambda(n)$. To $\sigma$ we associate a measure $\sigma_n$ on $[-1/2, 1/2]^d$ by setting

$$\sigma_n = \frac{1}{n^d} \sum_{x \in \Lambda(n)} \sigma(x)\, \delta_{x/n}$$

where $\delta_{x/n}$ is the Dirac mass at $x/n$. We call $\sigma_n$ the empirical magnetization. The expectation $b_n$ of $\sigma_n$ is

$$b_n = \frac{1}{n^{d+1}} \sum_{x \in \Lambda(n)} \sigma(x) x.$$

We denote by $\mathcal{L}^d$ or simply by $dx$ the $d$-dimensional Lebesgue measure.

**Theorem 3.1.** *Let $d \geq 2$ and let $T < T_c(d)$. There exists a bounded, closed, convex set $\mathcal{W}$ containing 0 in its interior, called the Wulff crystal of the Ising model such that the following holds.*

*Let $m < m^*$ be close enough to $m^*$ so that the rescaled Wulff crystal*

$$\mathcal{W}(m) = \left(\frac{m^* - m}{2m^*}\right)^{1/d} \frac{\mathcal{W}}{\mathcal{L}^d(\mathcal{W})^{1/d}}$$

*fits into the unit cube $[-1/2, 1/2]^d$. Let $w_n$ be the random measure defined by*

$$w_n(x)\, dx = \Big(1_{[-1/2,1/2]^d}(x) - 2 \cdot 1_{\mathcal{W}(m)}\Big(\frac{b_n}{m^* - m} + x\Big)\Big) m^*\, dx.$$

*This is the measure having density $-m^*$ on $-b_n/(m^* - m) + \mathcal{W}(m)$ and $m^*$ on the complement. Under the conditional probability*

$$\mu_n(\cdot) = \mu^+_{\Lambda(n),T}\Big(\,\cdot\,\Big|\frac{1}{n^d} \sum_{x \in \Lambda(n)} \sigma(x) \leq m\Big)$$

*the difference between the random measures $\sigma_n$ and $w_n$ converges weakly in probability towards* 0*, i.e., for any continuous function* $f : [-1/2, 1/2]^d \to \mathbb{R}$*,*

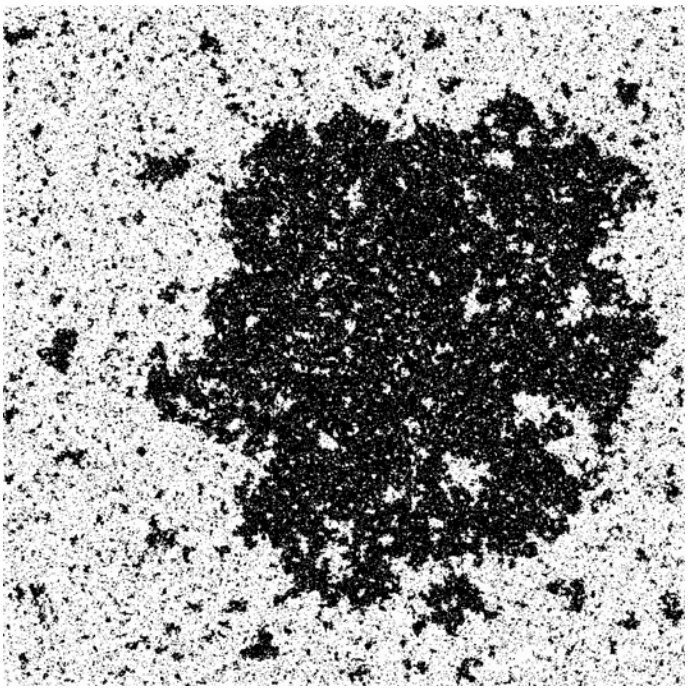
$$\lim_{n\to\infty} \mu_n\big(|\sigma_n(f) - w_n(f)| \geq \varepsilon\big) = 0 \quad \textit{for all } \varepsilon > 0.$$

*The probabilities of the deviations are of order* $\exp -cn^{d-1}$.

The last sentence of the theorem means the following. For any continuous function $f : [-1/2, 1/2]^d \to \mathbb{R}$, any $\varepsilon > 0$, there exist positive constants $b, c$ depending on $d$, $T$, $f$, $\varepsilon$ such that

$$\mu_n\Big(\Big|\frac{1}{n^d} \sum_{x\in\Lambda(n)} \sigma(x)\, f\Big(\frac{x}{n}\Big) + \int_{\mathcal{W}(m)} 2m^* f\Big(-\frac{b_n}{m^* - m} + x\Big)\, dx - \int_{[-1/2,1/2]^d} m^* f(x)\, dx\Big| > \varepsilon\Big) \leq b \exp(-cn^{d-1}).$$

The main assertion of the theorem is that the left-hand quantity goes to 0 for any continuous function $f$ and $\varepsilon > 0$. The objects appearing in the statement, namely the spontaneous magnetization $m^*$ and the Wulff crystal $\mathcal{W}$, are built as the thermodynamic limit of finite volume quantities. These objects can equivalently be defined with the help of the infinite volume Gibbs measure $\mu_T^+$.



Simulation of the Ising Wulff crystal at $T = 2.26$ after 69 days on a 1 Ghz PC.

Theorem 3.1 in dimension 2 is a consequence of the much finer results of Dobrushin, Kotecký, Shlosman [23] and Pfister [29] for low temperatures and Ioffe and Schonmann [25] for all subcritical temperatures. In dimensions 3 and higher, they were proven by Bodineau [7] for low temperatures and by Cerf and Pisztora [17] until the slab percolation threshold for temperatures such that the associated infinite volume FK measure is unique. Recently, Bodineau proved that this slab percolation threshold coincides with the true critical point [8] and that for any subcritical temperature, the associated infinite volume FK measure is indeed unique [9].

In two dimensions the Wulff droplet can be identified with a random region surrounded by a minus spin cluster. Its external boundary is therefore a large contour separating plus and minus spins which follows closely the boundary of the Wulff crystal in the sense of the Hausdorff metric [23], [25]. In dimension $d \geq 3$, it is widely believed that for low temperatures, the Wulff droplet can still be defined by a microscopic contour. However for temperatures close to $T_c$, a fundamentally new situation is expected. The dominant minus spin cluster of the Wulff droplet should percolate all the way to the boundary of the box. More precisely, there should exist two big spin clusters, one of pluses and one of minuses, and they should both be omnipresent in the entire box; the densities of these clusters should undergo an abrupt change at the boundary of the Wulff droplet. In this case the phase boundaries cannot be described directly with contours.

Let us mention the most recent works on the Wulff crystal. The low temperature expansion of the 3D Wulff crystal is computed in [14]. Alexander succeeded recently in deriving cube root fluctuations of the random curve around the Wulff crystal in the FK model in two dimensions [1], [4]. Couronné and Messikh provide a two dimensional version of Pisztora's coarse graining estimates [22]. Messikh analyzes the phase coexistence phenomenon in the 2D Ising model close to criticality: the Wulff crystal then becomes a circle [16], [27] (see also [28] for an application to image segmentation). Couronné has shown that the statistical repartition of the large finite clusters in the FK percolation model can be approximated by a Poisson process [21]. He has also studied the Wulff crystal for oriented percolation in dimension $d \geq 3$ [20]. In this model, the Wulff crystal has a singular point. Bodineau, Schonmann and Shlosman investigate the question of the flatness of the Wulff crystal [12]. Biskup, Chayes and Kotecky study the formation/dissolution of equilibrium droplets in the context of the 2D Ising model [5] and they derive the Gibbs–Thomson formula in the droplet formation regime [6]. Alexander, Biskup and Chayes devise an Ising-based model of a solvent-solute system [2], [3] and they study the associated phase separation phenomenon. The book [13] contains the proof of Theorem 3.1 and the corresponding statements in percolation.

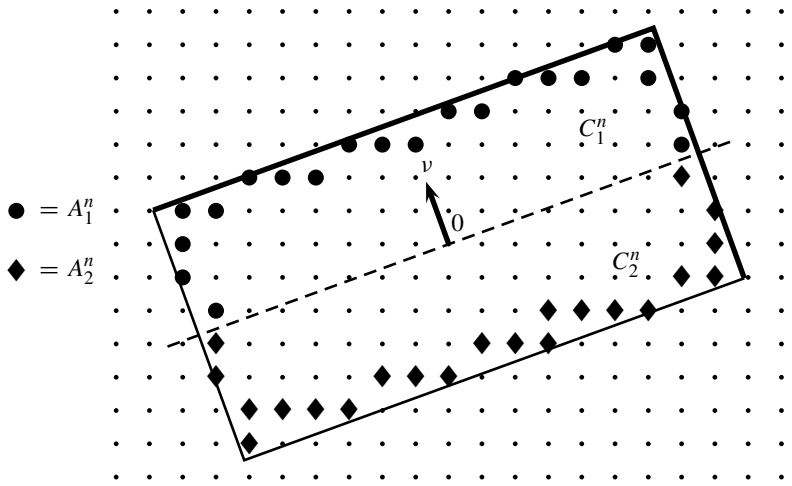
**The Wulff construction.** The Wulff crystal $\mathcal{W}$ appearing in Theorem 3.1 can be defined constructively. First we define the surface tension of the Ising model as follows. Let $\nu \in S^{d-1}$ be a unit vector in $\mathbb{R}^d$ and let $A$ be a unit hypersquare

orthogonal to $\nu$. Let $D_n$ be the cylinder

$$D_n = \{na + t\nu : a \in A,\ |t| \le n\}.$$

The set $D_n \setminus nA$ has two connected components, which we denote by $C_1^n$ and $C_2^n$. For $i = 1, 2$ and $n \in \mathbb{N}$, let $A_i^n$ be the set of the points of $C_i^n \cap \mathbb{Z}^d$ which have a nearest neighbour in $\mathbb{Z}^d \setminus D_n$:

$$A_i^n = \{x \in C_i^n \cap \mathbb{Z}^d : \text{there exists } y \in \mathbb{Z}^d \setminus D_n \text{ such that } |x - y| = 1\}.$$




The Ising Hamiltonian in $D_n$ is

$$H_n(\sigma) = -\frac{1}{2} \sum_{\substack{x,y \in D_n \\ |x-y|=1}} \sigma(x)\sigma(y) \quad \text{for all } \sigma \in \{-,+\}^{D_n}.$$

Let $\mathcal{E}_n$ be the set of the spin configurations $\sigma$ inside $D_n$ such that $\sigma(x) = +$ for $x \in A_1^n \cup A_2^n$. Let $\mathcal{F}_n$ be the set of the spin configurations $\sigma$ inside $D_n$ such that $\sigma(x) = -$ for $x \in A_1^n$ and $\sigma(x) = +$ for $x \in A_2^n$. The partition functions $Z_n^+$, $Z_n^{-,+}$ corresponding to pure $+$ and mixed $-, +$ boundary conditions at temperature $T$ are

$$Z_n^+ = \sum_{\sigma \in \mathcal{E}_n} \exp -\frac{H_n(\sigma)}{T}, \quad Z_n^{-,+} = \sum_{\sigma \in \mathcal{F}_n} \exp -\frac{H_n(\sigma)}{T}.$$

Let $T > 0$. The limit

$$\tau(\nu) = \lim_{n \to \infty} -\frac{1}{n^{d-1}} \ln \frac{Z_n^{-,+}}{Z_n^+}$$

exists in $[0, \infty]$. The function $\tau$ is called the surface tension of the Ising model. It satisfies the weak simplex inequality, it is continuous and invariant under the isometries which leave $\mathbb{Z}^d$ invariant. Moreover $\tau$ is positive in the regime $T < T_c$. The Wulff crystal $\mathcal{W}$ appearing in Theorem 3.1 is the Wulff shape associated to the surface tension $\tau$, called also the crystal of $\tau$, defined by

$$\mathcal{W} = \{x \in \mathbb{R}^d : x \cdot w \leq \tau(w) \text{ for all } w \text{ in } S^{d-1}\}.$$

**Large deviations.** The way to prove Theorem 3.1 is rather long. The key is the analysis of the deviations of the average magnetization from its typical value. We have first a weak law of large numbers:

$$\lim_{n\to\infty} \mu^+_{\Lambda(n),T}\Big(\frac{1}{n^d}\sum_{x\in\Lambda(n)}\sigma(x)\Big) = m^*(T) \quad \text{for all } T > 0.$$

The large deviations from above are similar in nature to what happens for a sum of independent identically distributed random variables.

**Theorem 3.2.** *Let $d \geq 2$ and let $T > 0$. For any $\alpha \in [-1, 1]$, the limit*

$$J(\alpha) = \lim_{n\to\infty} -\frac{1}{n^d} \ln \mu^+_{\Lambda(n),T}\Big(\frac{1}{n^d}\sum_{x\in\Lambda(n)}\sigma(x) \geq \alpha\Big)$$

*exists and is finite. The map $\alpha \in [-1, 1] \mapsto J(\alpha) \in \mathbb{R}^+$ is convex continuous. It vanishes on $[-1, m^*(T)]$ and it is strictly positive on $]m^*(T), 1]$.*

The deviations from above are of volume order. The function $J$ appearing in Theorem 3.2 vanishes on $[0, m^*[$ because on this interval the large deviations are of surface order. As we are in the phase coexistence regime, a new large deviation principle on the surface scale emerges.

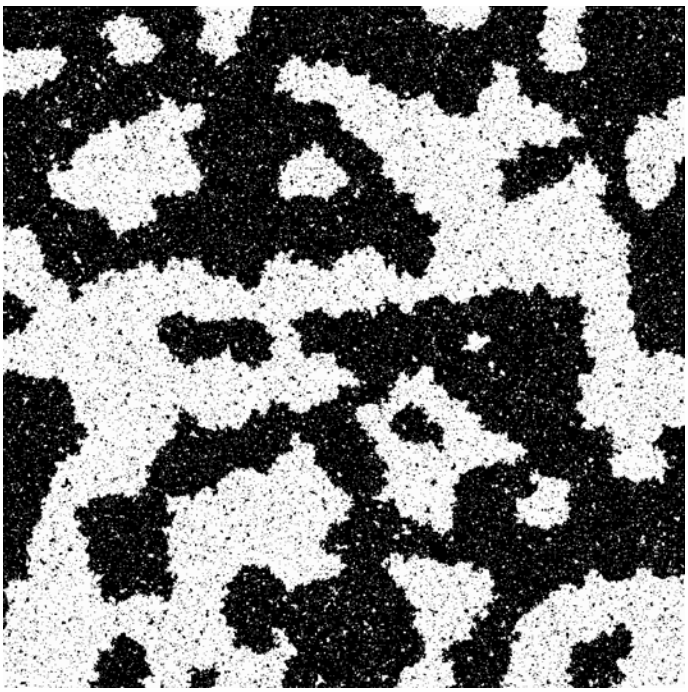
**Theorem 3.3.** *Let $d \geq 2$ and let $T < T_c(d)$. For $m < m^*$ close enough to $m^*$, so that the rescaled Wulff crystal $\mathcal{W}(m)$ fits into the unit cube $[-1/2, 1/2]^d$, we have*

$$\lim_{n\to\infty} \frac{1}{n^{d-1}} \ln \mu^+_{\Lambda(n),T}\Big(\frac{1}{n^d}\sum_{x\in\Lambda(n)}\sigma(x) \leq m\Big) = -d\left(\frac{m^* - m}{2m^*}\right)^{\frac{d-1}{d}} \mathcal{L}^d(\mathcal{W})^{\frac{1}{d}}.$$

Theorems 3.2 and 3.3 provide a complete picture of the large deviation behavior of the average magnetization. The emergence of a droplet is responsible for the asymptotic formula of Theorem 3.3. The full proofs of Theorems 3.2 and 3.3 can be found in the book [13].

## 4. Dynamics

Another very interesting topic is of course the dynamics. One should try to understand the dynamical mechanism leading to the creation of a Wulff crystal. It is naturally expected that several reasonable choices of microscopic dynamics induce the macroscopic dynamics associated to the motion by mean curvature. For instance it seems to be the case for the non-conservative Glauber dynamics, which is relevant for the beautiful theory of metastability (see [30], [31]). However the full understanding of these dynamics seems currently out of reach and only partial results are available [11], [15], [18], [19], [26], [32].



Phase separation under the Glauber dynamics at $T = 2.1$.

Let $d \geq 2$ and let $\Lambda \subset \mathbb{Z}^d$ be a cubic box. We wish to build a stochastic dynamics on the configuration space $\{-,+\}^\Lambda$ which models the microscopic repulsive forces between the particles. The interaction being microscopic, only one site or two neighbouring sites can be altered at each step. For $\sigma : \Lambda \to \{-,+\}$ and $x \in \Lambda$, we denote

$$S(\sigma, x) = \sum_{y:|x-y|=1} \sigma(y)$$

the sum of the spins of the neighbours of $x$.

Let $T > 0$ be a positive temperature. We consider two dynamics, which are built as discrete time Markov chains $(\sigma(k))_{k\geq 0}$ with state space $\{-,+\}^{\Lambda}$. We describe next the transition mechanisms of each dynamics.

**Glauber dynamics.** We suppose that $\sigma(k)$ is known and we explain how to build $\sigma(k+1)$. We first choose a site $x \in \Lambda$ randomly with the uniform law on $\Lambda$. We then compute $\Delta = 2\sigma(k,x)S(\sigma(k),x)$.

- If $\Delta < 0$, we flip the spin at $x$.
- If $\Delta \geq 0$, we flip the spin at $x$ with probability $\exp -(\Delta/T)$.

With the Glauber dynamics, at most one spin is changed at each time step.

**Kawasaki dynamics.** We suppose that $\sigma(k)$ is known and we explain how to build $\sigma(k+1)$. We first choose two neighbouring sites $x, y \in \Lambda$ randomly with the uniform law on the pairs of neighbours in $\Lambda$. We then compute $\Delta = (\sigma(k,x) - \sigma(k,y))(S(\sigma(k),x) - S(\sigma(k),y))$.

- If $\Delta < 0$, we exchange the spins of the sites at $x$ and $y$.
- If $\Delta \geq 0$, we exchange the spins of $x$ and $y$ with probability $\exp -(\Delta/T)$.

With the Kawasaki dynamics, at most two spins are changed at each time step.

A fundamental problem is to understand the scaling limits of these dynamics. For the Glauber dynamics, the adequate scaling is expected to converge to an anisotropic motion by mean curvature. More precisely, to the Markov chain $(\sigma(k))_{k\geq 0}$ we associate a process $(\sigma_n(t))_{t\geq 0}$ taking its values in the space of the Borel measures on $[-1/2,1/2]^d$ by setting

$$\sigma_n(t) = \frac{1}{n^d} \sum_{x\in\Lambda(n)} \sigma(\lfloor n^2 t \rfloor, x)\, \delta_{x/n} \quad \text{for all } t > 0,$$

where $\delta_{x/n}$ is the Dirac mass at $x/n$. We call $\sigma_n(t)$ the stochastic empirical magnetization.

Let $A$ be a subset of $[-1/2,1/2]^d$ having smooth boundary. We take as initial condition at step $n$ the configuration defined by

$$\sigma(0,x) = 1 - 2\cdot 1_{nA}(x) \quad \text{for all } x \in \mathbb{Z}^d.$$

Let $\mu$ be a function defined on the unit sphere $S^{d-1}$ of $\mathbb{R}^d$ with values in $\mathbb{R}^+$. Let $(A(t))_{t\geq 0}$ be the anisotropic mean curvature motion starting from $A$ associated to the function $\mu$, that is the solution (in some weak sense) of the equation:

$$v(x) = -\mu(\nu(x))\,\kappa(x)\,\nu(x) \quad \text{for all } t > 0,\ x \in \partial A(t),$$

where $v(x)$ is the speed at $x$, $\nu(x)$ is the normal vector to $A(t)$ at $x$ and $\kappa(x)$ is the curvature of $A(t)$ at $x$. Let $(w(t,x)\,dx)_{t\geq 0}$ be the measure valued process defined by

$$w(t,x)\,dx = m^*\big(1 - 2\cdot 1_{A(t)}(x)\big)\,dx \quad \text{for all } t \geq 0.$$

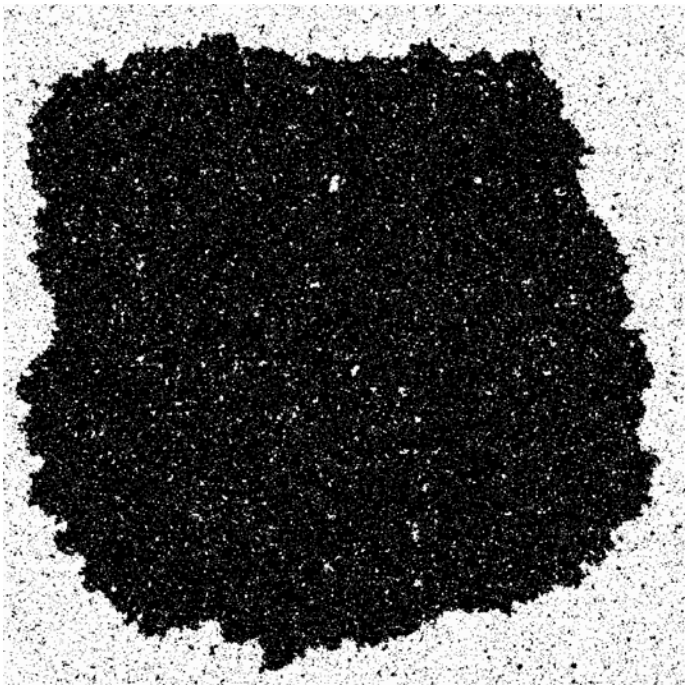
The big conjecture is that there exists a function $\mu\colon S^{d-1} \to \mathbb{R}^+$ such that the stochastic empirical magnetization $\sigma_n(t)$ converges weakly to the anisotropic mean curvature motion associated to $\mu$: there exists $T > 0$ such that, for any continuous function $f\colon [-1/2, 1/2]^d \to \mathbb{R}$,

$$\lim_{n\to\infty} P\big(|\sigma_n(t)(f) - w_n(t)(f)| \geq \varepsilon\big) = 0 \quad \text{for all } t \in [0,T],\ \varepsilon > 0.$$

We are still very far from the proof of such a result. In the two dimensional case at zero temperature, some specific computations [15], [32] show that, if this conjecture is correct, then the function $\mu$ governing the anisotropic motion by mean curvature must be equal to

$$\mu(x) = \frac{1}{2(|\cos\theta| + |\sin\theta|)^2},$$

where $\theta$ is the angle between the horizontal axis and the tangent to the boundary at $x$. The scaling limit of the conservative Kawasaki dynamics seems even more challenging to understand.



Evolution of a square droplet under Glauber dynamics at $T = 2.1$.

## References


[1] Alexander, K. S., Cube root boundary fluctuations for droplets in random cluster models. *Comm. Math. Phys.* **224** (3) (2001), 733–781.

[2] Alexander, K. S., Biskup, M., and Chayes, L., Colligative properties of solutions: I. Fixed concentrations. *J. Stat. Phys.* **119** (3–4) (2005), 479–507.

[3] Alexander, K. S., Biskup, M., and Chayes, L., Colligative properties of solutions: II. Vanishing concentrations. *J. Stat. Phys.* **119** (3–4) (2005), 509–537.

[4] Alexander, K. S., and Uzun, H., Lower bounds for boundary roughness for droplets in Bernoulli percolation. *Probab. Theory Relat. Fields* **127** (1) (2003), 62–88.

[5] Biskup, M., Chayes, L., and Kotecky, R., Critical region for droplet formation in the two-dimensional Ising model. *Comm. Math. Phys.* **242** (1–2) (2003), 137–183.

[6] Biskup, M., Chayes, L., and Kotecky, R., A proof of the Gibbs–Thomson formula in the droplet formation regime. *J. Stat. Phys.* **116** (1–4) (2004), 175–203.

[7] Bodineau, T., The Wulff construction in three and more dimensions. *Comm. Math. Phys.* **207** (1) (1999),197–229.

[8] Bodineau, T., Slab percolation for the Ising model. *Probab. Theory Relat. Fields* **132** (1) (2005), 83–118.

[9] Bodineau, T., Translation invariant Gibbs states for the Ising model. *Probab. Theory Relat. Fields* (2005), online first.

[10] Bodineau, T., Ioffe, D., and Velenik, Y., Rigorous probabilistic analysis of equilibrium crystal shapes. Probabilistic techniques in equilibrium and nonequilibrium statistical physics. *J. Math. Phys.* **41** (3) (2000), 1033–1098.

[11] Bodineau, T., and F. Martinelli, F., Some new results on the kinetic Ising model in a pure phase. *J. Stat. Phys.* **109** (1–2) (2002), 207–235.

[12] Bodineau, T., Schonmann, R. H., and Shlosman, S., 3D crystal: how flat its flat facets are? *Comm. Math. Phys.* **255** (3) (2005), 747–766.

[13] Cerf, R., *The Wulff crystal in Ising and percolation models*. Ecole d'été de Probabilités de Saint-Flour XXXIV (2004), Lecture Notes in Math. 1878, Springer-Verlag, Berlin 2006.

[14] Cerf, R., and Kenyon, R., The low temperature expansion of the Wulff crystal in the 3D Ising model. *Comm. Math. Phys.* **222** (1) (2001), 147–179.

[15] Cerf, R., and Louhichi, S., The initial drift of a 2D droplet at zero temperature. Preprint, 2005.

[16] Cerf, R., and Messikh, R. J., On the 2d Ising Wulff crystal near criticality. Preprint, 2006.

[17] Cerf, R., and Pisztora, A., On the Wulff crystal in the Ising model. *Ann. Probab.* **28** (3) (2000), 947–1017.

[18] Cesi, F., Guadagni, G., Martinelli, F., and Schonmann, R. H., On the 2D stochastic Ising model in the phase coexistence region near the critical point. *J. Stat. Phys.* **85** (1–2) (1996), 55–102.

[19] Chayes, L., Schonmann, R. H., and Swindle, G., Lifshitz' law for the volume of a two-dimensional droplet at zero temperature. *J. Statist. Phys.* **79** (5–6) (1995), 821–831.

[20] Couronné, O., The Wulff crystal for oriented percolation. Preprint, 2004.

[21] Couronné, O., Poisson approximation for large finite clusters in the supercritical FK model. *Markov Process. Related Fields*, to appear.

[22] Couronné, O., and Messikh, R. J., Surface order large deviations for 2D FK percolation and Potts models. *Stochastic Process. Appl.* **113** (1) (2004), 81–99.

[23] Dobrushin, R. L., Kotecký, R., and Shlosman, S. B., *Wulff construction: a global shape from local interaction*. Transl. Math. Monogr. 104, Amer. Math. Soc., Providence, RI, 1992.

[24] Ellis, R. S., *Entropy, Large Deviations and Statistical Mechanics*. Grundlehren Math. Wiss. 271, Springer-Verlag, New York 1985.

[25] Ioffe, D., and Schonmann, R. H., Dobrushin–Kotecký–Shlosman Theorem up to the critical temperature. *Comm. Math. Phys.* **199** (1) (1998), 117–167.

[26] Katsoulakis, M. A., Souganidis and Panagiotis, E., Generalized motion by mean curvature as a macroscopic limit of stochastic Ising models with long range interactions and Glauber dynamics. *Comm. Math. Phys.* **169** (1) (1995), 61–97.

[27] Messikh, R. J., On the surface tension of the 2d Ising model near criticality. In preparation.

[28] Messikh, R. J., Approximation of a Mumford–Shah functional using the Ising model: theory and numerics. In preparation.

[29] Pfister, C.-E., Large deviations and phase separation in the two-dimensional Ising model. *Helv. Phys. Acta* **64** (7) (1991), 953–1054.

[30] Schonmann, R. H., Slow droplet–driven relaxation of stochastic Ising models in the vicinity of the phase coexistence region. *Comm. Math. Phys.* **161** (1) (1994), 1–49.

[31] Schonmann, R. H., and Shlosman, S. B., Wulff droplets and the metastable relaxation of kinetic Ising models. *Comm. Math. Phys.* **194** (2) (1998), 389–462.

[32] Spohn, H., Interface motion in models with stochastic dynamics. *J. Statist. Phys.* **71** (5–6) (1993), 1081–1132.

[33] Wulff, G., Zur Frage der Geschwindigkeit des Wachstums und der Auflösung der Kristallflächen. *Z. Kristallogr.* **34** (1901), 449–530.

Département Mathématique, Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France

E-mail: rcerf@math.u-psud.fr

# Simple random covering, disconnection, late and favorite points

Amir Dembo

**Abstract.** We review recent advances in the study of the fractal nature of certain random sets, the key to which is a multi-scale truncated second moment method. We focus on some of the fine properties of the sample path of the most basic stochastic processes such as the simple random walk and the Brownian motion. As we shall see, probability on trees inspires many of our proofs, with trees used to model the relevant correlation structure. Along the way we also mention a few open problems.

**Mathematics Subject Classification (2000).** Primary 60J15; Secondary 28A80, 60G17, 60J65, 82C41.

**Keywords.** Random walk, Brownian motion, Gaussian free field, cover time, late points, favorite points, thick points, multi-fractal analysis, disconnection, intersection local time.

## 1. Introduction

The simple random walk (SRW) on a graph $G = (V, E)$ of finite degrees tracks the movement on the set $V$ of vertices by a particle which at each time step jumps with equal probability to any one of the nearest neighbors of its current position, independently of all previous positions. In particular, the SRW on $\mathbb{Z}^d$ is a fundamental object in probability theory. More than forty years ago, Erdős and Taylor posed in [ET60] the following problem about the SRW on $\mathbb{Z}^2$: What is the maximal number of visits by the walk to one lattice site during its first $n$ steps? More formally, denote by $L_n(x)$ the number of visits to $x$ by the SRW during its first $n$ steps, and set $L_n^* := \max_{x\in\mathbb{Z}^2} L_n(x)$. Then, it was conjectured in [ET60, (3.11)] and proved in [DPRZ01] that with probability one,

$$\lim_{n\to\infty} \frac{L_n^*}{(\log n)^2} = \frac{1}{\pi}. \tag{1.1}$$

As illustrated in Section 3, the key to proving (1.1) is a multi-scale truncated second moment method, inspired by the study of the corresponding problem for SRW on finite, regular trees. As detailed in Section 2, the same approach provides information about the location in $\mathbb{Z}^2$ where $L_n^*$ is attained, the number of sites $x \in \mathbb{Z}^2$ for which $L_n(x)$ is exceptionally large, and the fractal dimension of the corresponding object for the sample path of the planar Brownian motion.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The *cover time* $C_G$ for a SRW on a finite graph $G$ is the number of steps till the walk has visited all sites of $G$ at least once. It has been studied intensively by probabilists, statistical physicists, combinatorialists and computer scientists (e.g. [Ald89], [Bro90], [BH91], [NCF91], [MP94]). In particular, the SRW on a finite graph is a time-reversible Markov chain and the asymptotics of $C_G$ is an important aspect of the general theory of reversible Markov chains, see [AF01]. The problem of determining the asymptotics of the cover time $C_n := C_{\mathbb{Z}_n^2}$ for the two dimensional lattice torus $\mathbb{Z}_n^2 = \mathbb{Z}^2/n\mathbb{Z}^2$ of side length $n$ was posed by Wilf, see [Wil89], and more formally by Aldous, who conjectured in [Ald89] that

$$\lim_{n\to\infty} \frac{C_n}{(n\log n)^2} = \frac{4}{\pi} \quad \text{in probability,} \tag{1.2}$$

and proved the upper bound of $4/\pi$ in (1.2). A lower bound $2/\pi$ was later proved in [Law92] and the conjecture (1.2) resolved in [DPRZ04]. Whereas the general theory of reversible Markov chains provides the correct growth order of $C_n$ it fails to provide the multiplying constant in (1.2). As described in Section 4, the proof of (1.2) relies on the same multi-scale truncated second moment method used en-route to (1.1). Further, this approach provides information about the number and spatial distribution of the sites in $\mathbb{Z}_n^2$ for which the time of first visit by the SRW is of the order of the cover time. In addition, it allows us to answer the following questions of Révész about (discrete) discs covered by the random walk on $\mathbb{Z}^2$ till time $n$, namely, where every site of the lattice within the disc is visited by the walk at least once.

- What is the radius $\rho_n$ of the largest disc, centered at the origin that is covered during the first $n$ steps of the walk? It is shown in [DPRZ04] that

$$\lim_{n\to\infty} \mathbb{P}\Big( \frac{(\log \rho_n)^2}{\log n} \geq y \Big) = e^{-4y}, \tag{1.3}$$

  for all $y > 0$ (in 1989 Révész derived upper and lower bounds for (1.3) with non-matching constants 120 and $1/4$; these have been improved to 4 and 2 in [Law92] which also quotes (1.3) as a conjecture of Kesten).

- What is the radius $R_n$ of the largest disc (of arbitrary center) that is covered during the first $n$ steps of the walk? It is shown in [DPR07] that with probability one,

$$\lim_{n\to\infty} \frac{\log R_n}{\log n} = \frac{1}{4} \tag{1.4}$$

  (non-matching bounds for (1.4) with constants other than $1/4$ are given in [Rév93] where the existence of the limit is also conjectured).

In Section 5 we detail additional results for intersection local times and for the two dimensional Gaussian free field that have been obtained by the same approach we review here.

Statements such as (1.1) and (1.2) are easier to handle for the SRW on $\mathbb{Z}^d$, $d \geq 3$ whose transience enables us to effectively localize the relevant occupation measures, in contrast with the case of $d = 2$.

Let $G_n = \mathbb{Z}_n^d \times \mathbb{Z}$ denote the infinite discrete cylinder based on the $d$-dimensional discrete torus of side length $n$. We say that a finite subset $\Gamma$ disconnects $G_n$ if, for large $r$ the sets $\mathbb{Z}_n^d \times [r, \infty)$ and $\mathbb{Z}_n^d \times (-\infty, -r]$ are contained in two distinct connected components of $G_n \setminus \Gamma$. Consider the time $D_n$ till the range of the SRW on $G_n$ disconnects the cylinder. Clearly, the *disconnection time* $D_n$ is between the cover time $C_n$ of the base $\mathbb{Z}_n^d$ by the projection of the SRW and the cover time $\widehat{C}_n$ of the slice $\mathbb{Z}_n^d \times \{0\}$ by the SRW on $G_n$. It was shown in [DS06] that

$$\lim_{n\to\infty} \frac{\log D_n}{\log n} = \lim_{n\to\infty} \frac{\log \widehat{C}_n}{\log n} = 2d \quad \text{in probability.} \tag{1.5}$$

That is, the disconnection time is roughly of order $n^{2d}$ and comparable to $\widehat{C}_n$, but in contrast to the case of $d = 1$, when $d \geq 2$ it is substantially larger than the cover time $C_n$ (which is roughly of order $n^{\max(d,2)}$, up to logarithmic correction terms). We outline in Section 6 the geometric argument which is the key to the proof of (1.5) and explain in what sense (1.5) implies for $d \geq 2$ a massive clogging of the truncated cylinders of height $n^{d-\varepsilon}$ by the SRW before it disconnects the infinite cylinder. See also [Szn06] for recent universality results about the asymptotic of the disconnection time for the SRW on $H_n \times \mathbb{Z}$ (under mild conditions on the finite graph $H_n$).

For earlier surveys of parts of this body of work see [Per03], [Dem05], [Shi06]. Many additional interesting examples of random fractals as well as numerous references to earlier works on such problems are provided in the survey [Tay86]. See also [LeG92] for more about the planar Brownian path and [Lyo05] for probability on trees.

## 2. Favorite and thick points

**2.1. Favorite points for SRW on $\mathbb{Z}^d$.** Erdős and Révész in [ER84] call a site $x \in \mathbb{Z}^d$ for which $L_n(x) = L_n^*$ a *favorite point* of the SRW. In a similar manner, for any $0 < \alpha < 1$ we say that $x \in \mathbb{Z}^2$ is an *$\alpha$-favorite point* of the walk if $L_n(x) \geq (\alpha/\pi)(\log n)^2$. The size of the set $\mathcal{F}_n(\alpha)$ of $\alpha$-favorite points can be estimated by the same approach leading to (1.1). More precisely, it is shown in [DPRZ01] that for each $\alpha \in (0, 1]$,

$$\lim_{n\to\infty} \frac{\log |\mathcal{F}_n(\alpha)|}{\log n} = 1 - \alpha \quad \text{a.s.} \tag{2.1}$$

In other words, the $n^\beta$-most visited point during the first $n$ steps of the walk is visited approximately $\frac{1-\beta}{\pi}(\log n)^2$ times. This is in contrast with typical points on the path of the walk, each of which has order $\log n$ visits.

It is also shown in [DPRZ01] that any random sequence $\{x_n\}$ in $\mathbb{Z}^2$ such that $L_n(x_n)/L_n^* \to 1$ must satisfy

$$\lim_{n\to\infty} \frac{\log \|x_n\|}{\log n} = \frac{1}{2} \quad \text{a.s.} \tag{2.2}$$

In particular, the favorite points, i.e. those $x_n^*$ where $L_n^*$ is attained, are consistently located near the frontier of the set of visited points, at least on a logarithmic scale.

For the SRW on $\mathbb{Z}$ the analog of the statement (2.2) is contained in the results of Bass and Griffin [BG85]. See also [Rév05, page 160] for a list of unsolved problems about favorite points of the SRW on $\mathbb{Z}$, taken from [ER84]. Tóth provided recently a partial answer to one of these questions, showing that the SRW on $\mathbb{Z}$ has for sufficiently large time $n$ at most three different favorite points $x_n^*$ (cf. [Tót01] for this result and its history). However, not much is known about $x_n^*$. For example,

**Open Problem 2.1.**
• Determine for which $d \geq 1$ the SRW on $\mathbb{Z}^d$ has with probability one at most two favorite points $x_n^*$ for all $n$ sufficiently large.
• Describe the evolution of $n \mapsto x_n^*$ in any dimension $d \geq 2$.
• How is the growth of time between the first and last visits to $x_n^*$ prior to $n$, affected by the dimension $d$?

**2.2. Thick points for planar Brownian motion.** Let $w^{(m)}(t) = \sqrt{d/m}\, S_{\lfloor mt \rfloor}$ denote a time-space rescaled image of the SRW $(S_k, k \geq 0)$ on $\mathbb{Z}^d$. Donsker's functional CLT tells us that the distribution of $(w^{(m)}(t), t \geq 0)$ converges as $m \to \infty$, to that of the Brownian motion $(w(t), t \geq 0)$. The latter is a continuous in time, $\mathbb{R}^d$-valued Gaussian stochastic process, of independent coordinates, each starting at zero and having zero mean increments of variance $|t - s|$.

It is thus not surprising that the continuous time analogs of (1.1) and (2.1) can be expressed in terms of the Brownian *occupation measure*

$$\mu_t^w(A) = \int_0^t \mathbf{1}_A(w(s))ds, \quad \text{for all } A \subseteq \mathbb{R}^d \text{ Borel}, \tag{2.3}$$

in the planar case, that is, when $d = 2$. To this end, let $D(x, r)$ denote the open disc in $\mathbb{R}^2$, centered at $x$ and of radius $r$, and let $\bar{\theta} = \inf\{t \geq 0 : \|w(t)\| \geq 1\}$ be the exit time of the planar Brownian motion from the unit disc $D(0, 1)$. Since $w([0, \bar{\theta}]) := \{w(t) : 0 \leq t \leq \bar{\theta}\}$ is a compact set, it follows that $\mu_{\bar{\theta}}^w(D(x, r)) = 0$ for any $x \notin w([0, \bar{\theta}])$ and all $r$ small enough. Further, it is not hard to show that for almost all Brownian paths, the pointwise Hölder exponent of the random measure $\mu_{\bar{\theta}}^w$, namely,

$$\lim_{r\to 0} \frac{\log \mu_{\bar{\theta}}^w(D(x, r))}{\log r},$$

takes the same value 2 for all points $x \in w([0, \bar{\theta}])$ (see also [Ray63, Theorem 1] for the precise lim sup decay rate of $\mu_{\bar{\theta}}^w(D(0, r))$ when $r \to 0$). Therefore, standard

multi-fractal analysis must be refined in order to capture the delicate fluctuations of the Brownian occupation measure and obtain a non-degenerate dimension spectrum. Indeed, it is shown in [DPRZ01] that for any $0 < a \leq 2$,

$$\dim\{x : \lim_{r\to 0} \frac{\mu^w_{\bar\theta}(D(x,r))}{r^2(\log r)^2} = a\} = 2 - a \quad \text{a.s.} \tag{2.4}$$

(where throughout $\dim(A)$ denotes the Hausdorff dimension of the set $A$). For a typical $x$ on the Brownian path $\mu^w_{\bar\theta}(D(x,r)) \asymp r^2|\log r|$ (e.g. see [DPRZ01, Lemma 2.1]), so the *a-thick points*, i.e. those in the set considered in (2.4), correspond to unusually large occupation measure.

The identity (2.4), together with the appropriate upper bound, yields that

$$\lim_{r\to 0} \sup_{x\in\mathbb{R}^2} \frac{\mu^w_{\bar\theta}(D(x,r))}{r^2(\log r)^2} = 2 \quad \text{a.s.} \tag{2.5}$$

as conjectured by Perkins and Taylor. It is not hard to show then that both (2.4) and (2.5) hold when $\bar\theta$ is replaced by any deterministic $0 < T < \infty$ and it is in the latter form that (2.5) was stated as [PT87, Conjecture 2.4]. As shown in [DPRZ02, Theorem 1.2], both (2.4) and (2.5) apply even when the discs $D(x,r) = x + rD(0,1)$ are replaced by the sets $x + rK$, provided the set $K$ is normalized to have area (Lebesgue measure) $\pi$ and its boundary has zero Lebesgue measure.

See also [PPPY01] for the application of (2.5) to the problem of reconstructing the range of spatial Brownian motion from the occupation measure projected to the sphere.

**2.3. From Brownian motion to SRW.** The passage from (2.5) and (2.4) to the corresponding results (1.1) and (2.1) for the discrete setting is based on the celebrated strong approximation theorem of Komlós, Major and Tsunády [KMT75] which constructs in an enlarged probability space a one dimensional Brownian motion $w$ and a SRW $(S_k, k \geq 0)$ on $\mathbb{Z}$ such that $\mathbb{P}(\sup_{t\leq 1}|w(t) - w^{(n)}(t)| \geq c(\log n)/\sqrt{n}) \to 0$ when $n \to \infty$. A simple geometric argument extends this conclusion to the SRW on $\mathbb{Z}^2$. Further, applying Einmahl's multidimensional version of this strong approximation theorem, the same argument allows [DPRZ01, Theorem 5.1] to establish (1.1), (2.1) and (2.2) for a wide collection of two-dimensional lattice valued random walks whose increments are of zero mean and finite moments. Here is an outline of such an argument, revealing the source of the factor $2\pi$ between (1.1) and (2.5). Taking $r = r(n) = n^{\eta-1/2}$ for fixed $\eta > 0$ small, and fixing $0 < a < 2$, one predicts from (2.4) that there are about $r^{a-2}$ discs of radius $r$, which are $r$-separated of each other, each having a Brownian occupation measure of about $ar^2(\log r)^2$. By strong approximation a similar result applies for the occupation measure of $w^{(n)}$. Since $t \mapsto w^{(n)}(t)$ is piecewise constant on intervals of length $1/n$, this translates to $nar^2(\log r)^2$ visits by $w^{(n)}$ to discs whose radius is approximately $r$. Further, with each of these discs having about $\pi r^2 n/2$ of the sites of $\sqrt{2/n}\mathbb{Z}^2$, we see that to these discs correspond

distinct (random) points in $\mathbb{Z}^2$ having at least $2a(\log r)^2/\pi$ visits during the first $n$ steps of the SRW. So, for $\alpha = 2a(1/2-\eta)^2$ these $r^{a-2} = n^{1-2\eta-\alpha/(1-2\eta)}$ points are $\alpha$-favorite, and considering first $n \to \infty$ then $\eta \to 0$ gives the lower bound in (2.1) and consequently also in (1.1).

As shown in [Ros05], (2.1) and (1.1) can also be proved without reference to the Brownian motion results, by directly applying the multi-scale truncated second moment approach of Section 3 to the SRW on $\mathbb{Z}^2$.

## 3. The multi-scale truncated second moment

Fixing a positive integer $b \geq 2$ let $\Gamma_h$ denote the $b$-ary rooted regular tree of height $h$, that is, the degree of each vertex of $\Gamma_h$ is $b+1$, except for the root, denoted $\mathbf{o}$, whose degree is $b$ and the $b^h$ leaves of this tree, each of whom has degree one. Starting a SRW $(X_i)$ on $\Gamma_h$ at its left-most leaf let $\tau_{\mathbf{o}} = \inf\{i \geq 0 : X_i = \mathbf{o}\}$ denote the hitting time of the root of $\Gamma_h$ and $L_x$ the number of visits to $x \in \Gamma_h$ by $\{X_i, i \leq \tau_{\mathbf{o}}\}$. Fixing $0 < \alpha < 1$ we call a leaf $x$ of $\Gamma_h$ *$\alpha$-favorite* if $L_x \geq \alpha h^2 \log b$ and let $\mathcal{F}_h(\alpha)$ denote the set of $\alpha$-favorite leaves. In Subsection 3.1 we outline the multi-scale truncated second moment method in the context of proving that

$$\lim_{h\to\infty} \frac{1}{h} \log |\mathcal{F}_h(\alpha)| = (1-\alpha)\log b \quad \text{in probability.} \tag{3.1}$$

As demonstrated in Subsection 3.2, this is the core of the computations leading to (2.4) and (2.5), thereby also to (1.1), (2.1) and (2.2), where the trees $\Gamma_h$ serve in revealing the hidden correlation structure across scale (i.e., the radius of discs), and space (i.e., their centers).

**3.1. Favorite points for SRW on regular trees.** Let $\partial\Gamma_h$ denote the set of leaves of $\Gamma_h$ and $\mathbf{o} \leftrightarrow x$ denote the shortest path in $\Gamma_h$ between $\mathbf{o}$ and $x \in \partial\Gamma_h$, also called the ray of $x$. Fixing $x$ and projecting the SRW on its ray, we see that $L_x$ has the same law as the number of visits to a reflecting boundary at $h$ prior to absorption at 0, for a SRW $(Y_i)$ on $\{0, 1, \ldots, h\}$. In particular, $\mathbb{P}(L_x \geq t) \leq \left(1 - \frac{1}{h}\right)^{t-1}$ for any $x \in \partial\Gamma_h$, $t \geq 1$. Taking $t = t_h$ for $t_h := \alpha h^2 \log b$ yields the upper bound in (3.1) by an application of the first moment bound $\mathbb{P}(Z \geq 1) \leq \mathbb{E}Z$ for $Z = Z_h = b^{-\beta h}|\mathcal{F}_h(\alpha)|$ and $\beta > 1-\alpha$. With a little more work we find that as $h \to \infty$ also $\mathbb{E}(Z_h) \to \infty$ for $\beta < 1-\alpha$, supporting the validity of (3.1). To take advantage of diverging expectations one usually relies on the second moment method. That is, applying the classical bound $\mathbb{P}(Z \geq \delta \mathbb{E}Z) \geq (1-\delta)^2(\mathbb{E}Z)^2/\mathbb{E}Z^2$ for some $0 < \delta < 1$ and the preceding $Z = Z_h$. However, at least for $\alpha > 1/2$ this approach fails to work here. Indeed, as shown in [Dem05, Lemma 4.2], then $\mathbb{E}Z_h^2/(\mathbb{E}Z_h)^2 \to \infty$ since for such $\alpha$ and any $0 < \xi < 2\alpha - 1$ there exists $\eta = \eta(\alpha, \xi) > 0$ such that for large enough $h$

$$b^{-2(1-\alpha)h} \sum_{(x,y)\in\partial_\xi\Gamma_h} \mathbb{P}(L_x \geq t_h, L_y \geq t_h) \geq b^{\eta h}, \tag{3.2}$$

where $\partial_\xi \Gamma_h$ denote the collection of pairs $x, y \in \partial\Gamma_h$ for which $\mathbf{o} \leftrightarrow x$ and $\mathbf{o} \leftrightarrow y$ separate at distance $\xi h$ from $\mathbf{o}$.

Starting the SRW $\{Y_i\}$ at $Y_0 = h$ and assuming $t_h$ returns to $h$ prior to its absorption at 0, the expected number of excursions between $k-1$ and $k$ till the $t_h$-th return to $h$ is about $\alpha k^2 \log b$ when $k \gg 1$ and $h - k \gg 1$ (cf. [Dem05, Lemma 4.3]). This suggests that for a typical $x \in \mathcal{F}_h(\alpha)$ the SRW $(X_i)$ has prior to $\tau_{\mathbf{o}}$ about $\alpha k^2 \log b$ visits to a vertex on $\mathbf{o} \leftrightarrow x$ which is at distance $k$ from $\mathbf{o}$. The analysis leading to (3.2) reveals also that the main contribution of the collection $\partial_\xi \Gamma_h$ to the second moment of $|\mathcal{F}_h(\alpha)|$ is via the rare events of having prior to $\tau_{\mathbf{o}}$ a sufficiently excessive number of SRW excursions between the vertex $z$ where $\mathbf{o} \leftrightarrow x$ and $\mathbf{o} \leftrightarrow y$ separate and the leaves of the sub-tree rooted at $z$. The typical number of such excursions for $\alpha$-favorite leaves is only a fraction of what these events require, so they contribute little to $\mathbb{E}Z_h$ when $h$ is large. However, the occurrence of such rare event yields too many $\alpha$-favorite leaves at the sub-tree rooted at $z$, hence resulting with the excessive growth of the second moment of $Z_h$.

Since this problem occurs for any separation height $\xi h$, $\xi < 2\alpha - 1$, one should pursue a multi-scaling truncation strategy. That is, apply the second moment method for $Z_h = b^{-\beta h}|\mathcal{S}_h(\alpha)|$, replacing $\mathcal{F}_h(\alpha)$ by a subset $\mathcal{S}_h(\alpha)$ of leaves along the rays of which various excursion counts are kept within a relatively small distance from the typical excursion count profile for an $\alpha$-favorite leaf. Of course we are to do so while not changing much the mean of $Z_h$, that is, keeping

$$\mathbb{P}(x \in \mathcal{S}_h(\alpha)) = b^{-\alpha h(1+o(1))} \tag{3.3}$$

for all $x \in \partial\Gamma_h$. To attain (3.3) only $o(h)$ excursion counts are to be controlled along each ray of $\Gamma_h$. Specifically, fixing $c > 0$ large enough, we set heights $h_0 = 0$, $h_1 = 1$ and $h_k = [ck \log k]$ for $k = 2, \ldots, m$, taking $h = h_m$, and consider the number $N_k^x$ of complete excursions between vertices $x_{k-1}$ and $x_k$ at distances $h_{k-1}$ and $h_k$ from $\mathbf{o}$ along $\mathbf{o} \leftrightarrow x$ which occur between the first visit of the SRW to $x_1$ and its first successive visit to $x_0 = \mathbf{o}$. Recall that for a typical $\alpha$-favorite leaf $x$, the SRW $(X_i)$ makes about $\alpha h_k^2 \log b$ visits to $x_k$ during this time interval. With $\Delta_k = h_k - h_{k-1} = c \log k(1 + o(1))$ for large $k$, this translates to $N_k^x$ being near $n_k = \alpha h_k^2 \log b/(c \log k)$ for a typical $x \in \mathcal{F}_h(\alpha)$. Consequently, we take as $\mathcal{S}_h(\alpha)$ those $x \in \partial\Gamma_h$ such that $|N_k^x - n_k| \le k$ for $k = 2, \ldots, m$. Per $x \in \partial\Gamma_h$, the sequence $(N_k^x, k = 1, \ldots, m)$ is the realization of a non-homogeneous Markov chain on $\mathbb{Z}_+$, starting at $N_1^x = 1$. The transition probabilities of this chain are given by explicit hyper-geometric distributions. As $n_{k+1} = n_k(1 + 2/k)(1 + o(1))$, for large $k$ one gets by normal approximation that the probability of the transition from $N_k^x = [n_k]$ to $N_{k+1}^x = [n_{k+1}]$ is about $p_k = b^{-\alpha\Delta_k}/\sqrt{n_k}$. With $k/\sqrt{n_k} \to 0$, further analysis shows that for some finite constant $C$ which depends only on $(\alpha, b, c)$ and for any $|\ell_k - n_k| \le k$ and $|\ell_{k+1} - n_{k+1}| \le k + 1$, the probability of the transition from $N_k^x = \ell_k$ to $N_{k+1}^x = \ell_{k+1}$ is between $C^{-1}p_k$ and $Cp_k$ (cf. [Dem05, Lemma 4.6]). Note that $q_m = \mathbb{P}(x \in \mathcal{S}_{h_m}(\alpha))$ is the same for any $x \in \partial\Gamma_{h_m}$. Also, in the definition of the event $\{x \in \mathcal{S}_{h_m}(\alpha)\}$, for each $k$, the random variable $N_k^x$ can take any one of

$2k+1$ possible values. Hence, by the preceding analysis $m^{-1}\log\left(q_m/\left[\prod_{k=1}^{m-1} kp_k\right]\right)$ is bounded. Moreover, as $n_k = \zeta k^2 \log k$ for some positive constant $\zeta$, it follows that $kp_k = b^{-\alpha\Delta_k(1+o(1))}$ and with $m = o(h_m)$, it is now easy to verify that (3.3) holds (cf. [Dem05, Proposition 4.4]).

We turn next to study the correlation structure of $\{x \in \mathcal{S}_{h_m}(\alpha)\}$ across $x \in \partial\Gamma_{h_m}$. Specifically, let $B_{l,x} = \{|N_k^x - n_k| \le k, l < k \le m\}$ for $l = 1, \ldots, m-1$, noting that $B_{1,x} = \{x \in \mathcal{S}_{h_m}(\alpha)\}$, so the second moment of $|\mathcal{S}_{h_m}(\alpha)|$ is the sum of $\mathbb{P}(B_{1,x} \cap B_{1,y})$ over $x, y \in \partial\Gamma_{h_m}$. Let $q_{m,l}$ denote the maximum of these probabilities over pairs $(x, y)$ such that $x_1 = y_1$ and the rays $\mathbf{o} \leftrightarrow x$ and $\mathbf{o} \leftrightarrow y$ separate at a vertex $z$ whose distance from $\mathbf{o}$ is between $h_{l-1}$ and $h_l$. Since $B_{1,y} \subseteq B_{l,y}$, it thus suffices for us to get an upper bound on $\mathbb{P}(B_{1,x} \cap B_{l,y})$. To this end, note that given the value of $N_{l+1}^y$, the event $B_{l,y}$ is independent of $B_{1,x}$. Consequently,

$$q_{m,l} \le q_m \sum_{|\ell - n_{l+1}| \le l+1} \mathbb{P}(B_{l,y} | N_{l+1}^y = \ell).$$

By a similar reasoning,

$$q_m \ge q_{l+1} \inf_{|\ell - n_{l+1}| \le l+1} \mathbb{P}(B_{l,y} | N_{l+1}^y = \ell).$$

As we have already seen, the above terms $\mathbb{P}(B_{l,y}|N_{l+1}^y = \ell)$ are almost constant (up to a factor $C^2$) with respect to $\ell$, leading to the bound

$$q_{m,l} \le C^2(2l+3)\frac{q_m^2}{q_{l+1}} \tag{3.4}$$

(cf. [Dem05, Lemma 4.8]). This is effectively the same correlation structure as for independent percolation on the tree $\Gamma_{h_m}$ projected to skeleton heights $\{h_k\}$ and with level depending edge probabilities that are about $p_k$. In particular, using the bound of (3.4) leads to $\mathbb{E}[|\mathcal{S}_h(\alpha)|^2] \le K(\mathbb{E}|\mathcal{S}_h(\alpha)|)^2$ for some $K = K(\alpha)$ finite and any $h \in \{h_k\}_{k=1}^{\infty}$ (cf. [Dem05, Lemma 4.9]). Combining this with the first moment estimate of (3.3), an application of the second moment method yields that the probability of ${h_m}^{-1}\log|\mathcal{S}_{h_m}(\alpha)| \ge (1-\alpha)\log b(1+o(1))$ is bounded away from zero (being about $1/K$).

It remains to improve this result to one that holds with probability approaching one, and to connect the fact that $N_m^x$ is near $n_m$ with the event $\{x \in \mathcal{F}_{h_m}(\alpha)\}$. To this end, for $h \in [h_{m+1}, h_{m+2})$ and any vertex $v \in \Gamma_h$ of height $h - h_m + 1$ let $\mathcal{S}_{h_m}^v(\alpha)$ be defined in analogy to $\mathcal{S}_{h_m}(\alpha)$, but for the subtree rooted at the ancestor of $v$ and consisting of those $u \in \Gamma_h$ with $v$ on the shortest path from $u$ to $\mathbf{o}$. Next let $\mathcal{S}_h^*(\alpha)$ be the union of the sets $\mathcal{S}_{h_m}^v(\alpha)$ over the $R_h$ vertices $v$ of height $h - h_m + 1$ that the SRW on $\Gamma_h$ visits by time $\tau_{\mathbf{o}}$. The preceding bound applies for the SRW within any regular subtree of depth $h_m$, hence for each of the sets $\mathcal{S}_{h_m}^v(\alpha)$. Thus, as $R_h \to \infty$ with high probability, and the restrictions of the SRW to within the subtrees of $\Gamma_h$ rooted at these

vertices are independent of each other and of $R_h$, it follows that

$$\lim_{h\to\infty} \frac{1}{h} \log |\mathcal{S}_h^*(\alpha)| = (1-\alpha)\log b \quad \text{in probability}$$

(cf. [Dem05, Lemma 4.11]). Finally, a simple concentration argument shows that during $n_m - m$ excursions between $x \in \partial\Gamma_h$ and a vertex on its ray at distance $\Delta_m$ from $x$, the SRW visits $x$ less than $n_m\Delta_m(1-\delta)$ times with probability that decays to zero exponentially in $n_m$ (cf. [Dem05, Lemma 4.7]). This leads for any $\beta < \alpha$ to $\mathbb{P}(\mathcal{S}_h^*(\alpha) \subseteq \mathcal{F}_h(\beta)) \to 1$ and consequently completes the proof of (3.1).

**3.2. From trees to Brownian motion.** The approach of [DPRZ01] in proving (2.4) and (2.5), which goes back to [Ray63], is to control Brownian occupation measures using excursions between concentric discs. Specifically, fixing $R' > R > r$, the total occupation measure of $D(x, r)$ during the first $N$ excursions of the sample path between $D(x, R)$ and the complement of $D(x, R')$ is of the form $\sum_{i=1}^{N} \tau_i$, where $\tau_i$ denotes the occupation measure of $D(x, r)$ accumulated during the $i$-th such excursion. Since these Brownian excursions are independent of each other, so are the random variables $\{\tau_i\}$. Further, the events of interest here involve exceptionally large occupation measures that translate into having numerous excursions around the same point. Consequently, for the range of $N$ values relevant here, the resulting total occupation measure is highly concentrated around its mean $N\mathbb{E}\tau_1$.

From the corresponding elliptic PDE we have that $\mathbb{E}\tau_1 = r^2\log(R'/R)$ which by the strong Markov property of the Brownian motion implies also that $\mathbb{P}(\tau_1 \geq t[r^2\log(R'/r) + 1])$ decays exponentially in $t$ (cf. [Dem05, Lemma 5.5]). The statement (2.5) can be shown to concern the maximum of the occupation measures $\mu_{\bar{\theta}}^w(D(x_j, r))$ for a suitable non-random discrete net of about $r^{-2}$ points $x_j \in D(0, 1)$. To upper bound such maximum, we take $R' = 2$ in which case $\mu_{\bar{\theta}}^w(D(x, r)) \leq \tau_1$. Hence, for $a > 2$, we see from the tail probabilities of $\tau_1$ (at $t = a\log(1/r)$), that the expected number of discs with centers in this net and occupation measure exceeding $ar^2(\log r)^2$ decays to zero at rate $r^{a-2}$. By the first moment method we get the upper bound in (2.5), where the almost sure statement is attained by using the monotonicity of $r \mapsto \mu_{\bar{\theta}}^w((D(x, r))$ to interpolate between radii $r_n$ such that $\sum_n r_n^{a-2}$ is finite.

The same argument shows that the expected number of such discs with occupation measure exceeding $t_r = ar^2(\log r)^2$ diverges at rate $r^{a-2}$ when $a < 2$, supporting the validity of (2.5). Unfortunately, the second moment method fails to work here since the recurrence of the planar Brownian motion precludes a fast decay of the correlation between the events $\{\mu_{\bar{\theta}}^w((D(x, r)) \geq t_r\}$ even for $x$ and $y$ of distance $r^\delta$ of each other (and $\delta$ a fixed small positive constant). A natural truncation of the second moment is by localizing the occupation measures. For example, considering only the Brownian occupation measure of $D(x, r)$ during the interval between the first hitting time of this disc by the Brownian path and its successive exit of $D(x, R)$. For a suitable $R(r) \to 0$ such that $R(r)/r \to \infty$ the second moment of the number of discs with localized occupation measures exceeding $t_r$ is much smaller than the corresponding

second moment for $\mu_{\bar\theta}^w(D(x,r))$. However, it comes at the cost of a substantial drop in the corresponding first moment, so this strategy results with non-matching lower and upper bounds (cf. [PT87]).

We note in passing that the same problem arises in the context of the SRW on $\mathbb{Z}^2$, that is, in [ET60] treatment of (1.1), and more generally in all statements we make in this paper.

We mitigate this problem by employing a multi-scale truncation, as done in Subsection 3.1. To this end, recall that the counts of Brownian excursions between concentric discs of radii $\{e^{-j}\}$ have the same law as the bond occupation measure for the SRW on $\mathbb{Z}$, which plays a key role in proving (3.1). Indeed, for $x \notin D(0,r_1)$ taking $r_k = r_1 e^{-h_k}$ for $h_k = [ck\log k]$, $k = 2,\ldots,m$, the numbers $N_k^x$ of Brownian excursions from $\partial D(x,r_{k-1})$ to $\partial D(x,r_k)$ till hitting $\partial D(x,r_0)$ have the same joint law as the random variables $N_k^x$ for the SRW on the regular tree $\Gamma_{h_m}$ (to make this precise, take $r_0 = er_1$). Thus, setting $n_k = ah_k^2/(c\log k)$ and partitioning the square $S(1,1) = [2r_1, 4r_1]^2$ into $e^{2h_m}$ non-overlapping squares $S(m,i)$ of edge length $2r_m$ each, we now consider the random set $\mathcal{S}_{h_m}(a)$ that consists of centers $x = x_{m,i}$ of squares $S(m,i)$ for which $|N_k^x - n_k| \le k$, $k = 2,\ldots,m$. As explained above, the estimate (3.3) applies here (with $\alpha = a$, $b = e$ and $h = h_m$), resulting with $\mathbb{E}|\mathcal{S}_{h_m}(a)| = e^{(2-a)h_m(1+o(1))}$. We can further uniformly bound correlation terms of the form $\mathbb{P}(B_{1,x} \cap B_{l,y})$ for $x = x_{m,i}$ and $y = x_{m,j}$ such that $|x-y| \in (2r_l, 2r_{l-1}]$ in a similar manner to that for the SRW on regular trees, apart from two complications. First, to assure that $\partial D(x,r_k)$ does not intersect $\partial D(y,r_l)$ we have to exclude $k = l-1$. That is, remove from $B_{1,x}$ the constraints on the values of $N_{l-1}^x$ and $N_l^x$. Even after this is done, given $N_{l+1}^y = \ell$, the event $B_{l,y}$ still depends on $B_{1,x}$ via the locations of the initial and final points of the $\ell$ Brownian excursions from $\partial D(y,r_{l+1})$ to $\partial D(y,r_l)$. Using the Poisson kernel for the density of the exit location at $z \in \partial D(y,R')$ for a Brownian path starting at some $z' \in D(y,R')$, it can be shown that the probability of $B_{l,y}$ given $N_{l+1} = \ell$ *and the terminal points* of these $\ell$ excursions is at most $(1+\kappa r_{l+1}/r_l)^\ell \mathbb{P}(B_{l,y}|N_{l+1}^y = \ell)$ (cf. [DPRZ01, Lemma 7.4]). Taking $c$ sufficiently large ($c = 3$ will do), provides enough separation for $n_{l+1} \ll r_l/r_{l+1}$ resulting with a bound of the form of (3.4), apart from replacing $C^2(2l+3)$ by a larger polynomial factor (cf. [DPRZ01, Lemma 8.1]). This is enough for deducing that $\mathbb{E}[|\mathcal{S}_{h_m}(a)|^2] \le K(\mathbb{E}|\mathcal{S}_{h_m}(a)|)^2$, for some $K = K(a)$ and all $m$. Since $\mathbb{E}|\mathcal{S}_{h_m}(a)| \to \infty$, by the second moment method also $\liminf_m \mathbb{P}(\mathcal{S}_{h_m}(a) \neq \emptyset) \ge 1/K$. By Fatou's lemma, this implies the existence with positive probability of some random $m(j) \to \infty$ and $y_j \in \mathcal{S}_{h_{m(j)}}(a)$. By compactness of $S(1,1)$, the sequence $y_j$ has at least one limit point $y_* \in S(1,1)$.

Let $M_{x,r} = \mu_{\bar\theta}^w(D(x,r))/[r^2(\log r)^2]$. We claim that $M_{y_*,r} \to a$ for $r \to 0$. That is, $y_*$ is an $a$-thick point, see (2.4). To this end, recall that for $R' = r_{k-1}$ and $R = r = r_k$ we have $\mathbb{E}\tau_1 = r_k^2\Delta_k$, whereas $n_k\Delta_k$ is about $ah_k^2 = a(\log r_k)^2$. With $\{\tau_i\}$ of exponential tail probabilities, it follows that $\mathbb{P}(|M_{x,r_k} - a| > \eta, |N_k^x - n_k| \le k) \le e^{-\gamma n_k}$ for any $\eta > 0$, some $\gamma = \gamma(a,\eta) > 0$, all $k$ and $x \in S(1,1)$. By the monotonicity of the non-negative measure $\mu_{\bar\theta}^w$ and the almost sure continuity properties of $M_{x,r}$

in $x$ and $r$, we further deduce the existence of $\delta(r, w) \to 0$ such that with probability one, if $x \in S(1, 1)$ and $|N_k^x - k| \le n_k$ for all $k \le \ell$, then $|M_{x,r} - a| \le \delta(r)$ for all $r \ge r_\ell$ (cf. [DPRZ01, Section 6]). Recall that for any fixed $\ell$, if $j$ is sufficiently large then $|N_k^{y_j} - n_k| \le k$ for $k = 2, \ldots, \ell$. Consequently, $\limsup_j |M_{y_j,r} - a| \le \delta(r)$ for any $r > 0$. Since $|y_j - y_*| \to 0$, the monotonicity of $\mu_{\bar\theta}^w$ leads to $|M_{y_*,r} - a| \le \delta(r)$, and so we deduce that with some positive probability there exists an $a$-thick point.

Fixing $\beta < 2 - a$, by a slightly more involved argument, the squares $S(m, i)$ whose centers are in $\mathcal{S}_{h_m}(a)$ support the density with respect to Lebesgue measure of a random measure $\nu_m$ such that a non-zero weak limit point $\nu_\infty$ of $\{\nu_m\}$ is supported on a closed set of $a$-thick points and has a finite $\beta$-energy

$$\mathcal{E}_\beta(\nu_\infty) := \int |x - y|^{-\beta} d\nu_\infty(x) d\nu_\infty(y) \tag{3.5}$$

(cf. [DPRZ01, Section 3]). This in turns implies that with positive probability the dimension of the set of $a$-thick points is at least $\beta$. Using Brownian scaling, it follows by Blumenthal's zero-one law that the latter property holds with probability one, establishing the stated lower bound in (2.4) and hence also that of (2.5).

## 4. Fractal geometry: late points and covered discs

**4.1. Cover times and covered discs.** For $0 \le \gamma < 1$ let $C_n(\gamma)$ denote the time it takes until the largest disc unvisited by the SRW on the two dimensional lattice torus $\mathbb{Z}_n^2$ has radius $n^\gamma$. It is shown in [DPRZ04] that

$$\lim_{n\to\infty} \frac{C_n(\gamma)}{(n \log n)^2} = \frac{4}{\pi}(1 - \gamma)^2 \quad \text{in probability,} \tag{4.1}$$

with (1.2) corresponding to the special case of $\gamma = 0$. In other words, at time $\beta C_n$ the radius of the largest disc within the set unvisited by the SRW is $n^{(1-\sqrt{\beta})(1-o(1))}$.

In [PR04], Peres and Revelle deduce from (4.1) that the total variation convergence to the stationary measure for the SRW on the lamplighter group over $\mathbb{Z}_n^2$ requires at least $(4/\pi)(n \log n)^2(1 + o(1))$ steps. This is based on their observation that starting at a zero lamp configuration, the lamps remain identically zero on the sites of $\mathbb{Z}_n^2$ which the lamplighter did not visit, while under the stationary (uniform) measure of this SRW the probability of having a disc of radius $\log n$ in $\mathbb{Z}_n^2$ with a zero lamp configuration tends to zero with $n$. By general considerations, the asymptotic growth in $n$ of the total variation mixing time for the SRW on the lamplighter group over $\mathbb{Z}_n^d$, $d \ge 2$, is between $\mathbb{E}C_{\mathbb{Z}_n^d}$ and half its value, but it is not known whether this upper bound is tight also for $d \ge 3$ (see [PR04]).

Let $\{w_{\mathbb{T}^2}(t)\}$ denotes a Brownian motion on the two dimensional unit torus $\mathbb{T}^2$, with the corresponding hitting times, $\tau(x, \varepsilon) = \inf\{t > 0 : w_{\mathbb{T}^2}(t) \in D_{\mathbb{T}^2}(x, \varepsilon)\}$, and the $\varepsilon$-cover time,

$$\boldsymbol{C}_\varepsilon = \sup_{x\in\mathbb{T}^2} \{\tau(x, \varepsilon)\}. \tag{4.2}$$

Equivalently, $\boldsymbol{C}_r$ is the amount of time needed for the Wiener sausage of radius $r$ to completely cover $\mathbb{T}^2$. It is shown in [DPRZ04] that

$$\lim_{r\to 0} \frac{\boldsymbol{C}_r}{(\log r)^2} = \frac{2}{\pi} \quad \text{a.s.} \tag{4.3}$$

Similarly to the argument we outline in Subsection 2.3, (4.1) is an immediate consequence of (4.3). Indeed, we identify the vertices of $n^{-1}\mathbb{Z}_n^2$ with the corresponding subset of $\mathbb{T}^2$, noting that up to scaling by $n$, distances in $\mathbb{Z}_n^2$ match the corresponding distances in $\mathbb{T}^2$. Further identifying the latter with $[0,1)^2$, we represent $w_{\mathbb{T}^2}(t)$ as the image of the planar Brownian motion $w(t)$ via the non-expansive mapping $x \mapsto x \bmod \mathbb{Z}^2$. Similarly, if $S_k$ denotes the SRW on $\mathbb{Z}^2$, then $X_k = (n^{-1}S_k) \bmod \mathbb{Z}^2$ is the SRW on $n^{-1}\mathbb{Z}_n^2$. By Einmahl's [Ein89, Theorem 1] multidimensional version of KMT strong approximation theorem, we can construct $\{S_k\}$ and $\{w(t)\}$ on the same probability space such that with probability approaching one as $n \to \infty$, the distance in $\mathbb{T}^2$ between $w_{\mathbb{T}^2}(t)$ and $X_{\lfloor mt \rfloor}$ is for $m = 2n^2$ and $t \le 2(\log n)^2$ at most $\delta_n = (\log m)^3/\sqrt{m}$. Fixing $\gamma > 0$, since $\delta_n \ll n^{\gamma-1} = \varepsilon_n$ and $w_{\mathbb{T}^2}$ completely misses some disc $D_{\mathbb{T}^2}(x, \varepsilon_n)$ in $\mathbb{T}^2$ till time $\boldsymbol{C}_{\varepsilon_n}$, we deduce that $(X_k)$ completely misses a disc whose radius is about $\varepsilon_n$ till time $2n^2\boldsymbol{C}_{\varepsilon_n}$. That is, $C_n(\gamma)$ is about $2n^2\boldsymbol{C}_{\varepsilon_n}$. Taking $\gamma \to 0$ then provides the lower bound in (1.2). The matching upper bound on $C_n$ is easily obtained upon considering the expected number of sites that are unvisited during the first $\alpha(4/\pi)(n\log n)^2$ steps of the SRW on $\mathbb{Z}_n^2$, with $\alpha > 1$ fixed and $n \to \infty$.

The derivation of (1.3) follows a similar path, where setting $\varepsilon_m = m^{\gamma-1}$ for $\gamma > 0$ one first checks that the lower bound in (4.3) applies also for the $\varepsilon_m$-cover time of the disc $D_{\mathbb{T}^2}(0, br_m)$ when $r_m = c(\log m)^{-3}$ (with $b$, $c$ fixed and $m$ large). For $R > 0$ sufficiently small, the expected time it takes $w_{\mathbb{T}^2}$ to complete an excursion between the discs of radii $r_m$ and $R$, centered at the origin, is about $(1/\pi)\log(R/r_m)$. By a concentration argument the number of such excursions made by $w_{\mathbb{T}^2}$ till its $\varepsilon_m$-cover time of $D_{\mathbb{T}^2}(0, br_m)$ is thus about $(2/\pi)(\log \varepsilon_m)^2/[(1/\pi)(\log R/r_m)]$. For $1/2 < b < 1$ and $c < R/b$ this implies by a strong approximation argument that with high probability the planar SRW makes at least $(1-\gamma)^3 2(\log m)^2/(3\log\log m)$ excursions between the discs of radii $m(\log m)^3$ and $2m$, centered at the origin, till it covers the concentric disc of radius $m$ (cf. [DPRZ04, Lemma 5.1]). A similar argument leads to the matching upper bound. The limit distribution of (1.3) then follows upon studying the tail probabilities for the time it takes the planar SRW to complete one such excursion for large value of $m$ (cf. [Law92]). By a similar technique, [HP06] establishes the law of the iterated logarithm which corresponds to (1.3). That is,

$$\limsup_{n\to\infty} \frac{(\log \rho_n)^2}{\log n \log_3 n} = \frac{1}{4}, \quad \text{a.s.}$$

(where $\log_3$ denotes three iterations of the log function).

Jonasson and Schramm proved in [JS00] that if $G = (V, E)$ is a planar graph of maximal degree $D$ then there are constants $k_D$ and $K_D$ depending only on $D$ such

that $k_D|V|(\log|V|)^2 \le \mathbb{E}[C_G] \le K_D|V|^2$. As illustrated already, (1.2) is a direct consequence of (4.3), with the same argument applicable for finding the asymptotics of the cover time for SRW on different planar lattices. As explained in [DPRZ04, Section 9], this leads to

**Open Problem 4.1.** Is the lattice $\mathbb{Z}^2$ asymptotically the easiest to cover when $D = 4$? That is, does $\liminf \mathbb{E}[C_G]/(|V|(\log|V|)^2) = 1/\pi$, where the lim inf is over all planar graphs $G = (V, E)$ of maximal degree $D = 4$ and for $|V| \to \infty$?

The limit distribution of the cover time $C_n$ for the SRW on the lattice torus $\mathbb{Z}_n^2$ and similar random fluctuations are likely related to behavior such as that of branching Brownian motion, and hence to KPP-type partial differential equations (cf. [Ald91], [Bra83], [Bra86], [McK75]). However, very little is known about it. For example,

**Open Problem 4.2.** Does there exist a non-random sequence $b_n$ such that $b_n(\sqrt{C_n} - \mathrm{Med}\{\sqrt{C_n}\})$ converges in distribution to a non-degenerate random variable, and if so what is the limit distribution?

Even the existence of a non-random normalizing sequence $b_n$ that results with a tight, yet non-degenerate collection, is not obvious. See [BZ06] for a proof of tightness for the simpler problem in the context of SRW on the regular trees $\Gamma_h$.

For any $0 < \alpha < 1$, let $R_n(\alpha)$ denote the radius of the largest disc (of arbitrary center) consisting of $\alpha$-favorite points for the SRW on $\mathbb{Z}^2$. The proof of (1.4) combines the ideas behind the proofs of (1.1) and (1.2). Similarly, based on (2.1) and (4.1), it is also shown in [DPR07] that for any $0 < \alpha < 1$, with probability one,

$$\lim_{n\to\infty} \frac{\log R_n(\alpha)}{\log n} = \frac{1-\sqrt{\alpha}}{4}. \tag{4.4}$$

Indeed, with $h_k = [ck \log k]$ and $r_k = e^{h_k}$, it takes about $r_m^2$ steps for the SRW to first exit $D(0, r_m)$. Hence, taking $r_{m,k} = e^{h_m - h_k}$, (4.4) follows by showing that with probability one, if $\zeta = (b\beta - \sqrt{\alpha})/(1-\beta) > 1$ for some $b < 1$ and $m$ is large enough, there exists $x \in D(0, r_m)$ such that each $z \in D(x, r_{m,\beta m})$ is visited at least $(4\alpha/\pi)h_m^2$ times prior to the first exit of the SRW from $D(0, r_m)$, whereas no such $x$ exists if $\zeta < 1$ for some $b > 1$. To this end, for each $a < 2$, by strong approximation and the outline of the proof of the lower bound for (2.5), we have that for some $x \in D(0, r_m)$ and $n_k(a) = ah_k^2/(c \log k)$, prior to exiting $D(0, r_m)$ the SRW completes at least $n_k(a) - k$ excursions between $D(x, r_{m,k})$ and the complement of $D(x, r_{m,k-1})$. Thus, considering $k = \beta m - 1$ and $b < \sqrt{a/2}$ for which $\zeta = \zeta(b, \beta, \alpha) > 1$, for $m$ large enough and any $z \in D(x, r_{m,\beta m})$, the SRW completes at least $K = 2(b\beta)^2 h_m^2/(c \log m)$ excursions between $D(z, R)$ and the complement of $D(z, R')$ prior to exiting $D(0, r_m)$ (say, for $R = 2r_{m,\beta m-1}$ and $R' = 0.5r_{m,\beta m-2}$). Let $\widehat{\mathcal{L}}_K(\alpha)$ denote the collection of lattice sites $z \in D(x, r_{m,\beta m})$ visited less than $(4\alpha/\pi)h_m^2$ times during the first $K$ excursions of the SRW between $D(z, R)$ and the

complement of $D(z, R')$. Since there are only $\exp(2(1-\beta)h_m(1+o(1)))$ lattice sites in $D(x, r_{m,\beta m})$, upon showing that

$$\mathbb{P}(z \in \widehat{\mathcal{L}}_K(\alpha)) \le e^{-2(1-\beta)h_m\zeta^2(1+o(1))}, \tag{4.5}$$

it follows that $\mathbb{E}|\widehat{\mathcal{L}}_K(\alpha)| \to 0$ as $m \to \infty$ sufficiently fast to produce the lower bound in (4.4).

The bound (4.5) is obtained by a large deviations estimate of the following type. If $T_i$ are i.i.d. random variables, such that $\mathbb{P}(T_1 > t) \le Qe^{-t/M}$ for all $t > 0$, then for $\gamma < 1$ and $\lambda = (\gamma^{-1} - 1)/M > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{K} T_i \le \gamma^2 KQM\Big) \le e^{\lambda\gamma^2 KQM}\mathbb{E}[e^{-\lambda T_1}]^K \le e^{-(1-\gamma)^2 KQ}. \tag{4.6}$$

The $K$ excursions of the SRW between $D(z, R)$ and the complement of $D(z, R')$ are approximately independent of each other. Further, by potential theory estimates for the SRW (cf. [Law91]), the probability $Q$ that during such an excursion the SRW visits $z$ is about $(\log(R'/R))/\log R'$ which in turn is about $(c\log m)/((1-\beta)h_m)$. By similar reasoning, upon visiting $z$ at least once, the mean number of returns to $z$ by the SRW during such an excursion, denoted $M$ is about $\frac{2}{\pi}\log R'$, and the number $T_i$ of such returns to $z$ during the $i$-th excursion is such that $\mathbb{P}(T_i > t) \le Qe^{-t(1+o(1))/M}$ for all $t > 0$. Thus, with $KQ$ being about $2(b\beta)^2 h_m/(1-\beta)$ and $KQM$ about $(4(b\beta)^2/\pi)h_m^2$, setting $\gamma = \sqrt{\alpha}/(b\beta) < 1$ in (4.6) leads to (4.5).

As for the matching upper bound in (4.4), by the preceding reasoning, for each $a > 2$, with probability that is fast approaching one (in $m$) for every $x \in D(0, r_m)$ the SRW completes fewer than $n_k(a)$ excursions between $D(x, r_{m,k})$ and the complement of $D(x, r_{m,k-1})$ by the time it exits $D(0, r_m)$. Further, if $\zeta < 1$ for some $b > 1$ then for $k = \beta m - 1$ the expected size of the set $\widehat{\mathcal{L}}_K(\alpha)$ of $z \in D(x, r_{m,k+1})$ with less than $(4\alpha/\pi)h_m^2$ visits by the SRW during its first $K = n_k(2b^2)$ excursions between $D(x, r_{m,k})$ and the complement of $D(x, r_{m,k-1})$ diverges as $m \to \infty$. A truncated multi-scale second moment argument similar to that of Subsection 4.2 then allows us to deduce that for each fixed $x$, with probability approaching one (in $m$), the set $\widehat{\mathcal{L}}_K(\alpha)$ is non-empty, thus completing the proof of (4.4).

**4.2. Late points for SRW on regular trees.** The cover time problem (4.3) is, in a sense, dual of (2.5), in that it replaces "extremely large" occupation measure by "extremely small" occupation measure. Indeed, the derivation in [DPRZ04] of the lower bound for (4.3) is based on another toy problem involving the SRW on a regular tree. In this case it is the asymptotics of the number $\widetilde{C}_h$ of returns to $\mathbf{o}$ by the SRW on $\Gamma_h$, starting at $\mathbf{o}$, till it visits all leaves of the tree.

Both [Ald91] and [Per03] show that $\mathbb{E}\widetilde{C}_h = h^2 b(1+o(1))\log b$ as $h \to \infty$. However, these proofs rely on an embedded branching process argument exploiting the tree structure of $\Gamma_h$ and as such are not suitable to deal with the corresponding

Brownian result (4.3). We explain next how the multi-scale truncated second-moment provides another derivation of the asymptotic of $\widetilde{C}_h$ which is *robust enough* to be adapted in [DPRZ04] to deal with the Brownian motion setting of (4.3).

Turning to deal with $\widetilde{C}_h$, fixing $\alpha > 0$ we say that a leaf $x$ of $\Gamma_h$ is $\alpha$-*late* if the number $\widetilde{R}_x$ of returns to $\mathbf{o}$ by the SRW on $\Gamma_h$ till its first visit to $x \in \partial\Gamma_h$ is at least $\alpha h^2 b \log b$. Starting at $\mathbf{o}$, the probability that the SRW visits a specific leaf $x$ before returning to $\mathbf{o}$ is $1/(bh)$. Hence $\mathbb{P}(\widetilde{R}_x \geq t) = \left(1 - \frac{1}{bh}\right)^t$, yielding the first moment estimate

$$\mathbb{E}(|\mathcal{L}_h(\alpha)|) = b^{h(1-\alpha)(1+o(1))},$$

for the set $\mathcal{L}_h(\alpha)$ of $\alpha$-late leaves of $\Gamma_h$. By the first moment method, this shows that for any $\alpha > 1$ the set $\mathcal{L}_h(\alpha)$ is empty with high probability. Since $\widetilde{C}_h$ is the maximum of $\widetilde{R}_x$ over $x \in \partial\Gamma_h$, we deduce that $\widetilde{C}_h$ is about $h^2 b \log b$ upon showing that for each $0 < \alpha < 1$,

$$\lim_{h\to\infty} \frac{1}{h} \log |\mathcal{L}_h(\alpha)| = (1-\alpha)\log b \quad \text{in probability.} \tag{4.7}$$

As in the derivation of the asymptotic (3.1) for the number of $\alpha$-favorite leaves, the second moment of $|\mathcal{L}_h(\alpha)|$ is much larger than $b^{2h(1-\alpha)(1+o(1))}$. Let $h_k = [ck \log k]$ and $n_k = \alpha h_k^2 \log b/(c \log k)$. Then, adapting the approach taken in Subsection 3.1, the appropriate way of truncating $\mathcal{L}_h(\alpha)$ is by considering for $h \in [h_{m+\rho}, h_{m+\rho+1})$ and $v \in \Gamma_h$ of height $h - h_m + 1$, the subset $\mathcal{S}^v_{h_m}(\alpha)$ of leaves $x$ in the subtree rooted at $v$ such that $N_2^x = 0$ and $|N_k^x - n_k| \leq k$ for $k = 3, \ldots, m-1$. In duality with the case of $\alpha$-favorite leaves, here we set $x_1 = x$ and for $k = 2, \ldots, m$ the vertex $x_k$ is at distance $h_k - 1$ from the *leaf* $x$ along the ray $v \leftrightarrow x$ (so that now $x_m = v$). We then have $N_k^x$ count the number of complete excursions between $x_{k-1}$ and $x_k$ which occur during the first $n_m$ excursions of the SRW between vertices $x_{m-1}$ and $x_m$.

Omitting the detailed computations for this case, we note in passing that similarly to the derivation of (3.3), here $\bar{q}_m = \mathbb{P}(x \in \mathcal{S}^v_{h_m}(\alpha)) = b^{-\alpha h_m(1+o(1))}$, yielding the desired asymptotic growth of $\mathbb{E}(|\mathcal{S}^v_{h_m}(\alpha)|)$. Further, similarly to the derivation of (3.4), here we have that

$$\bar{q}_{m,l} := \sup_{x_{l-1} \neq y_{l-1}} \mathbb{P}\big(x \in \mathcal{S}^v_{h_m}(\alpha) \text{ and } y \in \mathcal{S}^v_{h_m}(\alpha)\big) \leq C l \bar{q}_m \bar{q}_{l-1},$$

for some $C < \infty$ and $l = 2, \ldots, m$ (with the convention of $\bar{q}_1 = \bar{q}_2 = 1$). Although $m \mapsto \mathbb{E}(|\mathcal{S}^v_{h_m}(\alpha)|^2)/\mathbb{E}(|\mathcal{S}^v_{h_m}(\alpha)|)^2$ is now unbounded, its polynomial growth is easily accommodated by considering the sum $|\mathcal{S}^*_h(\alpha)|$ of the $b^{h-h_m+1}$ i.i.d. random variables $|\mathcal{S}^v_{h_m}(\alpha)|$, provided $\rho$ is large enough.

To complete the derivation of (4.7) it thus remains only to show that with high probability $\mathcal{S}^*_h(\alpha) \subseteq \mathcal{L}_h(\beta)$ for $\beta < \alpha$. To this end, recall that the condition $N_2^x = 0$ guarantees that $x \in \mathcal{S}^v_{h_m}(\alpha)$ is not visited by the SRW during its first $n_m$ excursions between the vertex $v$ at level $h - h_m + 1$ and a specific descendent $u = u(x)$ of $v$ which is $\Delta_m$ levels further from $\mathbf{o}$. A concentration argument similar to that of

[Dem05, Lemma 4.7] shows that the probability that the number of returns to $\mathbf{o}$ during the first $n_m$ excursions between such pair $v$ and $u$ is less than $(1-\delta)n_m b\Delta_m$, decays exponentially in $n_m$. Since $n_m b\Delta_m = \alpha b h^2(1+o(1))\log b$ it follows that $\mathbb{P}(\mathcal{S}_h^*(\alpha) \subseteq \mathcal{L}_h(\beta)) \to 1$ for any $\beta < \alpha$ and $h \to \infty$, as required.

**4.3. Clustering of late points on $\mathbb{Z}_n^2$.** Simulations of the SRW on $\mathbb{Z}_n^2$ reveal that the points that are visited late by the walk appear in clumps of various sizes. Motivated by [BH91], the geometric characteristics of these clumps are studied in [DPRZ06]. More precisely, with $\tau_x$ denoting the first hitting time of $x$ by the SRW on $\mathbb{Z}_n^2$, it is shown in [DPRZ06] that for $\alpha \in (0,1]$ the set

$$\mathcal{L}_n(\alpha) = \{x \in \mathbb{Z}_n^2 : \ \tau_x \geq \alpha(4/\pi)(n\log n)^2\},$$

of *$\alpha$-late points* for the SRW has typical size $n^{2(1-\alpha)+o(1)}$ and that for any fixed $x$ in $\mathbb{Z}_n^2$ and $0 < \beta < 1$,

$$\lim_{n\to\infty} \frac{\log|\mathcal{L}_n(\alpha) \cap D(x, n^\beta)|}{\log n} = 2\beta - 2\alpha/\beta \quad \text{in probability} \tag{4.8}$$

(with $D(x, n^\beta)$ the disc of radius $n^\beta$, centered at $x \in \mathbb{Z}_n^2$). If the points of $\mathcal{L}_n(\alpha)$ were approximately evenly spread out in $\mathbb{Z}_n^2$, then the number of $\alpha$-late points in $D(x, n^\beta)$ would be $n^{2\beta-2\alpha+o(1)}$, whereas (4.8) shows that there are significantly less of them (as $2\beta - 2\alpha/\beta < 2\beta - 2\alpha$). The clustering pattern of $\mathcal{L}_n(\alpha)$ is confirmed by another result of [DPRZ06], showing that for any $0 < \alpha, \beta < 1$, if $Y_n$ is chosen uniformly in $\mathcal{L}_n(\alpha)$ then,

$$\lim_{n\to\infty} \frac{\log|\mathcal{L}_n(\alpha) \cap D(Y_n, n^\beta)|}{\log n} = 2\beta(1-\alpha) \quad \text{in probability.} \tag{4.9}$$

Counting pairs of late points one is tempted to apply the approximations

$$\begin{aligned}
&\mathbb{E}[\#\text{ of pairs of } \alpha\text{-late points within distance } n^\beta \text{ of each other}] \\
&\quad \simeq n^2 n^{2\beta}\mathbb{P}(x,\ y \ \text{ are } \alpha\text{-late when } |x-y| \simeq n^\beta) \\
&\quad \simeq (\text{Typical value of such number of pairs}) \\
&\quad \simeq (\#\text{ discs in } n^\beta \text{ grid with } \alpha\text{-late points}) \\
&\qquad \times (\text{Typical value of } |\mathcal{L}_n(\alpha) \cap D(x, n^\beta)| \text{ when } x \text{ is } \alpha\text{-late})^2.
\end{aligned} \tag{4.10}$$

However, as seen in [DPRZ06], such approximations fail to hold and these three quantities exhibit different power growth exponents.

All preceding results about the clustering of $\alpha$-late points are derived in [DPRZ06] by the same truncated multi-scale second moment method as in Subsection 4.2. In contrast to the results of [DPRZ04] about cover times and covered discs, the derivation of (4.9) requires conditioning upon the first hitting time at a *point* $x \in \mathbb{Z}_n^2$ for which strong approximation theorems are ineffective. This is handled in [DPRZ06] by

appealing to potential theory estimates for SRW (cf. [Law91]) and relying on the fact that only rough approximations of the probability that $|N_k^x - n_k| \le k$ are required. The derivation of the power growth exponent for (4.10) is more technically challenging since the mean of this object is already off its typical value. Further, the dominant contribution is from pairs of $\alpha$-late points having significantly less excursions between discs at the intermediate scale $n^\beta$ (in comparison with the typical excursion count profile for $\alpha$-late points), forcing an accumulation of many $\alpha$-late points inside such a disc. Thus, the evaluation of this power growth exponent is by a large deviations analysis of various excursion counts, similar in spirit to that of (4.5).

Not much else is known about the late points. For example,

**Open Problem 4.3.**
• In [DPRZ06] the power growth exponent of pairs of $\alpha$-late points within distance $n^\beta$ of each other is computed. Extend this to a "full multi-fractal analysis". For example, find the power growth exponent of triplets $(x_1, x_2, x_3)$ of $\alpha$-late points, such that $x_i$ is within distance $n^\beta$ of $x_j$ for $i, j = 1, 2, 3$.
• What is the distribution of the distance between the last two points to be covered by the SRW in $\mathbb{Z}_n^2$? In particular, does the chance that they are adjacent go to zero as $n$ grows?

**Open Problem 4.4.** Adapting the proof of (4.3) one can show that for any $a \le 2$,

$$\dim\left\{x \in \mathbb{T}^2 : \limsup_{\varepsilon\to 0} \frac{\tau(x,\varepsilon)}{(\log\varepsilon)^2} = \frac{a}{\pi}\right\} = 2 - a \quad \text{a.s.} \tag{4.11}$$

It is not clear what to do when the lim sup in (4.11) is replaced by a limit or a lim inf, since in this case we can no longer avoid considering the stochastic behavior of $\tau(x, \varepsilon)$ across different scales (i.e. $\varepsilon$ values) which are highly dependent when the scales are close to each other.

## 5. Intersection local times and Gaussian free fields

**5.1. Intersections and processes with jumps.** While having the Markov property is of much help for the results described here, [DPRZ02] deals with thick points for intersection of planar sample path, whereby it is partially lost. For example, [DPRZ02, Theorem 1.4] provides the analog of (2.4), showing that for any $0 < a \le 1$,

$$\dim\left\{x : \lim_{r\to 0} \frac{\mathcal{I}_{\bar\theta,\bar\theta'}(D(x,r))}{r^2(\log r)^4} = a^2\right\} = 2 - 2a \quad \text{a.s.} \tag{5.1}$$

where $\mathcal{I}_{T,T'}(A)$ denotes the projected intersection local time of two independent planar Brownian motions $(w(t), 0 \le t \le T)$ and $(w'(t'), 0 \le t' \le T')$, normalized by factor $\pi$ (see [LeG92, Chapter VIII] for more on $\mathcal{I}_{T,T'}(\cdot)$ and its properties). By strong approximation this leads to the analogs of (1.1) and (2.1) for the intersections of

two independent SRW on $\mathbb{Z}^2$ (cf. [DPRZ02, Theorem 1.1]). The lower bound in (5.1) is proved by first constructing for $\beta < 2-a$, along the lines of Subsection 3.2, a non-zero random measure $\nu'_\infty$ of finite $\beta$-energy (cf. (3.5)), that is supported on a closed set of $a$-thick points for $w'([0,\bar\theta'])$. The same construction is then repeated for $w([0,\bar\theta])$, now using the squares $S(m,i)$ whose centers are in $\mathcal{S}_{h_m}(a)$ to define the density of the random measure $\nu_m$ with respect to $\nu'_\infty$, instead of with respect to Lebesgue measure. Fixing $\gamma < \beta - a < 2-2a$, the non-zero weak limit point of $\{\nu_m\}$ is then of finite $\gamma$-energy and shown in [DPRZ02, Section 4.1] to be supported on a closed subset of the *$a^2$-thick intersection points* of (5.1). This strategy works since $\mathcal{I}_{\bar\theta,\bar\theta'}(A)$ is a continuous additive functional for $w([0,\bar\theta])$ with Revuz measure $\pi\rho$ such that $\rho(B) = \mu^{w'}_{\bar\theta'}(A\cap B)$; for almost every path $w'([0,\bar\theta'])$, the accumulation on $D(x,r)$ of such an additive functional during one excursion of $w(t)$ between $D(x,R)$ and the complement of $D(x,R')$ has a mean value $\rho(D(x,r))\log(R'/R) \pm Cr^2(\log r)^2$ and exponentially decaying tail probabilities (cf. [DPRZ02, Lemma 2.3]).

With a slightly different approach of directly controlling the excursion counts for two random walks, [DPR07, Theorem 1.3] shows that the radius of the largest discrete disc in the intersection of the sample path of two independent SRW on $\mathbb{Z}^2$, each run for $n$ steps, is $R_{n,2} = n^{1/(2+2\sqrt{2})+o(1)}$. However,

**Open Problem 5.1.** The growth rate of the diameter $D_{n,2}$ of the largest connected component of the intersection of two independent planar simple random walk path, each run for $n$ steps, is not known. A related open problem is to determine whether the intersection of two independent planar Brownian motion path, each run for a unit time, is almost surely a totally disconnected set.

The results described here depend mostly on the local properties of the stochastic processes considered. They are thus not limited to Brownian motion or to random walks. In particular, sample path continuity is not essential. For instance, Daviaud [Dav05] considers the Cauchy process on $\mathbb{R}$, that is, a stochastic process $X(t)$ with $X(0) = 0$ and stationary independent increments $X(t+s) - X(t)$ each of whom has the Cauchy density $s/(\pi(s^2+x^2))$ with respect to Lebesgue measure. This is a recurrent process, whose Green's function has a logarithmic behavior, similar to that of the planar Brownian motion (cf. [Dav05, Proposition 1.3]), but which has infinitely many jumps. In analogy with (2.4), [Dav05] shows that for $0 \le a \le 1$,

$$\dim\{x : \lim_{r\to 0} \frac{\mu^X_{\bar\theta}(D(x,r))}{r(\log r)^2} = \frac{2}{\pi}a\} = 1-a \quad \text{a.s.}$$

where $D(x,r)$ is an interval of radius $r$ and center $x \in \mathbb{R}$ and $\bar\theta = \inf\{t : |X(t)| \ge 1\}$. To avoid the technical difficulties due to jumps, [Dav05] relies on the representation of the Cauchy process $X(t)$, up to a well understood time change, as the intersection of the planar Brownian motion $w(t)$ and, say, the $x$-axis, thereby adapting the strategy of [DPRZ02] to the case at hand.

**5.2. The Gaussian free field.** The discrete $d$-dimensional Gaussian free field (abbreviated GFF) on the square $V_n = \{1, \dots, n\}^d \subset \mathbb{Z}^d$ is a Gaussian random vector $(\phi_x,\ x \in V_n)$ of zero mean and covariance given by the Green's function of the SRW restricted to $V_n$. That is, $\mathbb{E}[\phi_x \phi_y]$ is the expected number of visits to $y$ by the SRW on $\mathbb{Z}^d$, starting at $x$ and run till its first exit from $V_n$. The GFF (also called the harmonic crystal) is a special case of the *solid on solid* model used in statistical physics to describe the effective interface between two phases at low temperature (cf. [Gia01], [Fun05] and the references therein).

In this context, the presence of a hard wall is manifested by the *entropic repulsion* conditioning on the non-negativity constraint $\Omega^+_{n,\varepsilon} = \{\phi_x \geq 0 : x \in V_{n,\varepsilon}\}$, for small $\varepsilon > 0$, where $V_{n,\varepsilon}$ denotes the subset of points in $V_n$ of distance at least $n\varepsilon$ to the boundary of $V_n$. For $d \geq 3$ it is well known that entropic repulsion pushes the GFF far from the wall, while asymptotically making no other changes to its law (cf. [Gia01, Chapter 3] and the references therein). Both [BDG01] and [Dav06] consider the effect of entropic repulsion on the two dimensional GFF. To this end, [BDG01, Theorem 2] shows that $\max_{x \in V_n} \phi_x$ grows with $n$ like $g_n = 2\sqrt{2/\pi} \log n$, from which they deduce that upon conditioning on $\Omega^+_{n,\varepsilon}$, the GFF is shifted by $g_n$, that is, $g_n^{-1}\phi_x \to 1$ uniformly on $V_{n,\varepsilon}$.

Theorem 2 of [BDG01] is derived by a multi-scale truncated second moment approach which is motivated by the similarity between the GFF and a branching random walk type model on regular trees. Whereas the profile of excursion counts along each ray is the key object of study in Section 3.1, the approach of [BDG01] is to consider a notion of success shared by all vertices of $\Gamma_h$ at a given height (i.e. distance from $\mathbf{o}$). Adapted to the context of (3.1), the $h_k$-th level of $\Gamma_h$ is successful if for enough vertices $x_k$ of $\Gamma_h$ at height $h_k$, the SRW completed by time $\tau_{\mathbf{o}}$ at least $n_k - k$ excursions between $x_k$ and its ancestor $x_{k-1}$ (at height $h_{k-1}$). Starting at distance $\delta h$ from $\mathbf{o}$ with enough independence to have sufficiently many such vertices at distance $h_2$ further from $\mathbf{o}$, the success of the $h_k$-th level propagates (in $k$) by counting the excursions to vertices at height $h_{k+1}$ during the first $n_k - k$ excursions between their ancestors at heights $h_{k-1}$ and $h_k$. By controlling the union over the probabilities of failure at the different steps $k = 2, 3, \dots, m$ of this process, one concludes that with high probability, the last step, consisting of $\partial\Gamma_h$, is successful.

Pursuing the same approach, [Dav06] goes further in relating the conditioned GFF with the shifted GFF by deriving in this context the analogs of the results of [DPRZ06] and [DPR07]. For example, it is shown in [Dav06, Theorem 1.1] that conditioned on $\Omega^+_{n,\varepsilon}$, the largest disc within $V_{n,\varepsilon}$ for which all values of $\phi_x$ are below $\eta g_n$ is of radius $n^{\eta/2(1+o(1))}$. Indeed, for the unconditional GFF this is the radius of the largest disc for which all values of $\phi_x$ exceed $(1-\eta)g_n$ (see [Dav06, Theorem 1.7]), a result which is the analog of (4.4). The transformation $\eta = \sqrt{\alpha}$ between the two has to do with the isomorphism between $\phi_x^2/2$ and the local time at $x$ of a continuous time planar SRW.

**Open Problem 5.2.** The results of [Dav06] suggest the possibility of simpler proofs in the SRW world by proving the corresponding results for the GFF and applying an isomorphism theorem. Can you find such a proof for (1.1)?

We note in passing that "level lines" of the continuous two dimensional GFF are intimately related to the conformally invariant Schramm–Loewner evolution (abbreviated SLE). See [She06] for a survey of the continuous GFF, [Wer04] for the SLE and its application for computing intersection exponents of independent Brownian motions and [SS06] for the convergence of the zero level interface of an interpolated GFF with appropriate boundary values to variants of the SLE(4) process. However, though the path of the planar Brownian motion is also conformally invariant, we do not know of any direct relation between the results presented here and the SLE.

## 6. Disconnection of cylinders by random walks

Let $(X_k)$ denote the SRW on the infinite discrete cylinder $G_n = \mathbb{Z}_n^d \times \mathbb{Z}$ (endowed with its natural graph structure), which starts at $X_0 = 0$. As $X_k$ is an irreducible, recurrent Markov chain, it is easy to see that the disconnection time $D_n = \inf\{k \geq 0 : \ X_{[0,k]}$ disconnects $G_n\}$ is almost surely finite. Further, $C_n \leq D_n \leq \widehat{C}_n$, where $C_n$ denotes the first time the projection of $X_\cdot$ has visited all points of the base $\mathbb{Z}_n^d$ and $\widehat{C}_n$ denotes the cover time of the slice $\mathbb{Z}_n^d \times \{0\}$ by the SRW on $G_n$. When $d = 1$, it is straightforward to argue that $D_n$ is roughly of order $n^2$ and comparable to $C_n$ (and to $\widehat{C}_n$). Indeed, by the general theory of Markov chains one knows that for any $d \geq 1$, the sequence $\log C_n / \log n$ converges in probability to $\max(d, 2)$ (as we have seen already, much more in known about $C_n$). From (1.5) we see a different behavior for $d \geq 2$ in which case $D_n = n^{2d+o(1)}$ is much larger than $C_n$.

The upper bound in (1.5) is quite simple to prove. It is based on the fact that $D_n \leq \widehat{C}_n$ and a relatively crude upper bound on the cover time $\widehat{C}_n$ of the slice $\mathbb{Z}_n^d \times \{0\}$. Indeed, fixing $\beta > d - 1$, though the hitting times of the sites on this slice are of infinite mean, the first moment method works for the number $Z_n$ of non-visited sites on it during the first $n^\beta$ excursions of the SRW between the truncated cylinders $\mathbb{Z}_n^d \times [-n, n]$ and $\mathbb{Z}_n^d \times [-2n, 2n]$. For large $n$ it gives with high probability an upper bound on $\widehat{C}_n$ by the time it takes the walk to make these $n^\beta$ excursions, which in turn is bounded above by $n^\gamma$ for fixed $\gamma > 2(\beta + 1)$.

Somewhat surprisingly, this rather primitive strategy of replacing $D_n$ by $\widehat{C}_n$ captures the correct rough order of magnitude of $D_n$. However, the lower bound is more delicate because a direct enumeration over the huge collection of possible disconnecting subsets of $G_n$ seems to lead nowhere. Instead, we find a robust geometric property that every disconnecting $\Gamma \subseteq G_n$ must have, then show that with high probability $X_{[0,n^{2d-\delta}]}$ lacks this property. More precisely, [DS06, Lemma 2.4] shows that for $\gamma \in (0, 1)$ fixed and any $n$ large enough, if $\Gamma$ disconnects $G_n$ then there exists a box of side length $n^\gamma$ in $G_n$ that contains at least order of $n^{d\gamma}$ points of $\Gamma$. This

is a purely combinatorial argument, based on the following isoperimetric inequality [DP96, (A.3)]:

For any $\varepsilon > 0$ and finite box $B \subset \mathbb{Z}^{d+1}$,

$$|A \cap B| \le (1-\varepsilon)|B| \implies |\partial_B(A \cap B)| \ge \delta |A \cap B|^{d/(d+1)}, \tag{6.1}$$

where $\partial_B(U)$ denotes the points of $B \setminus U$ within distance one of $U$ and the positive constant $\delta$ depends only on $\varepsilon$ and $d \ge 1$. As shown in [DP96], the inequality (6.1) is a direct consequence of the Loomis–Whitney inequality bounding the size of any finite set $A \subset \mathbb{Z}^{d+1}$ by the $d$-th root of the product of sizes of the $(d+1)$ projections of $A$ on the hyperplanes perpendicular to the coordinate axes (cf. [LW49, Theorem 2]).

The preceding combinatorial argument is complemented by a probabilistic analysis involving the excursions between two concentric boxes of side length $n^\gamma$ and $2n^\gamma$. Fixing $1 > \delta > 3(d-1)\gamma > 0$, it shows that for some finite $c_0$ the probability that during its first $n^{2d-\delta}$ steps the SRW makes more than $c_0 \log n$ such excursions for some pair of boxes, decays to zero as $n \to \infty$. Scaling the occupation measure of the smaller box during one such excursion by $n^{-2\gamma}$ yields a random variable whose moment generating function is bounded uniformly in $n$ and the excursion's starting point. Consequently, for some finite $c_1$ depending only on $\delta$ and $\gamma$, during its first $n^{2d-\delta}$ steps, the number of visits by the SRW to any box of side length $n^\gamma$ does not exceed $c_1(\log n)\, n^{2\gamma}$. For $d \ge 3$ and large $n$, this is substantially less than the order of $n^{d\gamma}$ points that the walk must have visited for at least one such box by time $D_n$ (since by definition $X_{[0,D_n]}$ disconnects $G_n$), producing in this case the lower bound on $D_n$ as stated in (1.5).

Though the argument for $d = 2$ is of a similar flavor, it requires a considerable refinement in order to utilize the much smaller differences, of only a logarithmic growth in $n$, that we have here. To this end, by similar isoperimetric controls, [DS06, Lemma 2.5] shows that for some finite, positive constants $c_i$, $i = 2, 3, 4$, and $n$ large enough, if $\Gamma$ disconnects $G_n$, then for one of the three two-dimensional coordinate projections there exists a box of side length $n^\gamma$ and a collection of $c_2(\log n)^{2\alpha}$ disjoint sub-boxes, each of side length $\ell_n = n^\gamma (\log n)^{-\alpha}$, whose centers lie on a common $c_3\ell_n$-sub-grid of this box, such that the projection of the intersection of $\Gamma$ with any of these sub-boxes, contains at least $c_4 \ell_n^2$ points. The stated lower bound on $D_n$ in case $d = 2$ is thus the result of a more careful probabilistic analysis which shows that for $\gamma$ small and $\alpha < 3/4$ the probability that the set $\Gamma = X_{[0,n^{4-\delta}]}$ has this property, tends to zero as $n \to \infty$.

One consequence of (1.5) is that when $d \ge 2$ and $n$ is large, by the time $D_n$ the walk pretty much fills up the truncated cylinders of height $n^{d-\varepsilon}$. More precisely, with $\rho(x, A)$ denoting the minimal length of a nearest neighbor path from $x \in G_n$ to $A \subseteq G_n$, for any $d \ge 2$, $\varepsilon > 0$ and $\eta > 0$,

$$\lim_{n\to\infty} n^{-\eta} \max_{x \in \mathbb{Z}_n^d \times [-n^{d-\varepsilon}, n^{d-\varepsilon}]} \rho(x, X_{[0,D_n]}) = 0 \quad \text{in probability.} \tag{6.2}$$

This is in contrast with the situation when $d = 1$, where with non-vanishing prob-

ability there are points in such a truncated cylinder which are at distance $n$ from $X_{[0,D_n]}$. The clogging effect of (6.2) is a direct consequence of the lower bound $n^{2d-\delta}$ on $D_n$, due to (1.5), as one can show that within $n^{2d-\delta}$ steps, in a uniform fashion for $\mathbb{Z}_n^d \times [-n^{d-\varepsilon}, n^{d-\varepsilon}]$, the walk comes "often enough" within distance $n$ of $x$, giving it each time an opportunity to come even closer to $x$.

**Acknowledgment.** I thank Yuval Peres, Jay Rosen and Ofer Zeitouni, for valuable feedback on a preliminary version of this manuscript, and the National Science Foundation for funding the research on which it is based.

## References

[AF01] Aldous, D. J., and Fill, J., Reversible Markov chains and random walks on graphs. http://stat-www.berkeley.edu/users/aldous/RWG/book.html, 2001.

[Ald89] Aldous, D. J., *Probability approximations via the Poisson clumping heuristic*. Appl. Math. Sci. 77, Springer-Verlag, New York 1989.

[Ald91] Aldous, D., Random walk covering of some special trees. *J. Math. Analysis Appl.* **157** (1) (1991), 271–283.

[BG85] Bass, R. F., and Griffin, P. S., The most visited site of Brownian motion and simple random walk. *Z. Wahrsch. Verw. Gebiete* **70** (1985), 417–436.

[BDG01] Bolthausen, E., Deuschel, J.-D., and Giacomin, G., Entropic repulsion and the maximum of the two dimensional free field. *Ann. Probab.* **29** (4) (2001), 1670–1692.

[Bra83] Bramson, M., *Convergence of solutions of the Kolmogorov equation to travelling waves*. Mem. Amer. Math. Soc. 44, no. 285, Amer. Math. Soc., Providence, RI, 1983.

[Bra86] Bramson, M., Location of the traveling wave for the Kolmogorov equation. *Probab. Theory and Related Fields* **73** (4) (1986), 481–515.

[BZ06] Bramson, M., and O. Zeitouni, O., Recursions and tightness. Preprint, 2006.

[Bro90] Broder, A., Universal sequences and graph cover times. A short survey. In *Sequences* (Naples/Positano, 1988), Springer-Verlag, New York 1990, 109–122.

[BH91] Brummelhuis, M., and Hilhorst, H., Covering of a finite lattice by a random walk. *Physica A* **176** (1991), 387–408.

[Dav05] Daviaud, O., Thick points for the Cauchy process. *Ann. Inst. H. Poincaré Probab. Statist.* **41** (2005), 953–970.

[Dav06] Daviaud, O., Extremes of the discrete two-dimensional gaussian free field. *Ann. Probab.* **34** (3) (2006).

[Dem05] Dembo, A., Favorite points, cover times and fractals. In *École d'été de probabilitiés de Saint-Flour XXXIII – 2003*, Lecture Notes in Math. 1869, Springer-Verlag, Berlin 2005, 5–108.

[DPR07] Dembo, A., Peres, Y., and Rosen, J., How large a disc is covered by a random walk in $n$ steps? *Ann. Probab.* **35** (2007).

[DPRZ01] Dembo, A., Peres, Y., Rosen, J., and Zeitouni, O., Thick points for planar Brownian motion and the Erdős-Taylor conjecture on random walk. *Acta Math.* **186** (2001), 239–270.

[DPRZ02] Dembo, A., Peres, Y., Rosen, J., and Zeitouni, O., Thick points for intersections of planar Brownian paths. *Trans. Amer. Math. Soc.* **354** (2002), 4969–5003.

[DPRZ04] Dembo, A., Peres, Y., Rosen, J., and Zeitouni, O., Cover times for Brownian motion and random walks in two dimensions. *Ann. of Math.* **160** (2004), 433–464.

[DPRZ06] Dembo, A., Peres, Y., Rosen, J., and Zeitouni, O., Late points for random walks in two dimensions. *Ann. Probab.* **34** (1) (2006), 219–263.

[DS06] Dembo, A., and Sznitman, A., On the disconnection of a discrete cylinder by a random walk. *Probab. Theory Related Fields* (2006).

[DP96] Deuschel, J.-D., and Pisztora, A., Surface order large deviations for high-density percolation. *Probab. Theory Related Fields* **104** (4) (1996), 467–482.

[Ein89] Einmahl, U., Extensions of results of Komlós, Major and Tusnády to the multivariate case. *J. Multivariate Anal.* **28** (1989), 20–68.

[ER84] Erdős, P., and Révész, P., On the favorite points of a random walk. In *Mathematical Structures - Computational Mathematics - Mathematical Modelling* 2, Publ. House Bulgar. Acad. Sci., Sofia 1984, 152–157.

[ET60] Erdős, P., and Taylor, S. J., Some problems concerning the structure of random walk paths. *Acta Math. Acad. Sci. Hungar.* **11** (1960), 137–162.

[Fun05] Funaki, T., Stochastic interface models. In *École d'été de probabilitiés de Saint-Flour XXXIII – 2003*, Lecture Notes in Math. 1869, Springer-Verlag, Berlin 2005, 109–274.

[Gia01] Giacomin, G., *Aspects of statistical mechanics of random surfaces*. Lecture notes, IHP 2001.

[HP06] Hough, J. B., and Peres, Y., An LIL for cover times of disks by planar random walk and Wiener sausage. *Trans. Amer. Math. Soc.* **358** (2006).

[JS00] Jonasson, J., and Schramm, O., On the cover time of planar graphs. *Electron. Comm. Probab.* **5** (10) (2000), 85–90.

[KMT75] Komlós, J., Major, P., and Tusnády, G., An approximation of partial sums of independent RVs, and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** (1975), 111–131.

[Law91] Lawler, G., *Intersections of random walks*. Probab. Appl., Birkhäuser, Boston, MA, 1991.

[Law92] Lawler, G., On the covering time of a disc by a random walk in two dimensions. In *Seminar in Stochastic Processes* 1992, Birkhäuser, Basel 1993, 189–208.

[LeG92] LeGall, J.-F., Some properties of planar Brownian motion. In *École d'été de probabilités de Saint-Flour XX – 1990*, Lecture Notes in Math. 1527, Springer-Verlag, Berlin 1992, 111–235.

[LW49] Loomis, L. H., and Whitney, H., An inequality related to the isoperimetric inequality. *Bull. Amer. Math. Soc.* **55** (1949), 961–962.

[Lyo05] Lyons, R., Probability on trees and networks. http://mypage.iu.edu/~ rdlyons/ prbtree/book.pdf (with Y. Peres), 1997-2005.

[McK75] McKean, H. P., Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piskunov. *Comm. Pure Appl. Math.* **28** (1975), 323–331.

[MP94] Mihail, M., and Papadimitriou, C. H., On the random walk method for protocol testing. In *Computer aided verification* (Stanford, CA, 1994), Lecture Notes in Comput. Sci. 818, Springer-Verlag, Berlin 1994, 132–141.

[NCF91] Nemirovsky, A. M., and Coutinho-Filho, M. D., Lattice covering time in D dimensions: theory and mean field approximation. *Physica A* **177** (1991), 233–240.

[Per03] Peres, Y., Brownian intersections, cover times and thick points via trees. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing 2002, 73–78.

[PPPY01] Pemantle, R., Peres, Y., Pitman, J., and Yor, M., Where did the Brownian particle go? *Electron. Comm. Probab.* **6** (10) (2001).

[PR04] Peres, Y., and Revelle, D., Mixing times for random walks on finite lamplighter groups. *Electron. Comm. Probab.* **9** (2004), 825–846.

[PT87] Perkins, E. A., and Taylor, S. J., Uniform measure results for the image of subsets under Brownian motion. *Probab. Theory Related Fields* **76** (1987), 257–289.

[Ray63] Ray, D., Sojourn times and the exact Hausdorff measure of the sample path for planar Brownian motion. *Trans. Amer. Math. Soc.* **106** (1963), 436–444.

[Rév93] Révész, P., Clusters of a random walk on the plane. *Ann. Probab.* **21** (1993), 318–328.

[Rév05] Révész, P., *Random walk in random and non-random environments* Second edition, World Scientific, Singapore 2005.

[Ros05] Rosen, J., A random walk proof of the Erdős-Taylor conjecture. *Period. Math. Hungar.* **50** (2005), 223–245.

[She06] Sheffield, S., Gaussian free fields for mathematicians. *Probab. Theory Related Fields* (2006).

[Shi06] Shi, Z., *Problèmes de recouvrement et points exceptionnels pour la marche aléatoire et le mouvement Brownian*. Séminaire Bourbaki, exposé n$^o$ 951, 2004/05, 2006.

[SS06] Schramm, O., and Sheffield, S., Contour lines of the two-dimensional discrete gaussian free field. Preprint, 2006.

[Szn06] Sznitman, A., How universal are asymptotics of disconnection times in discrete cylinders? Preprint, 2006.

[Tay86] Taylor, J., The measure theory of random fractals. *Math. Proc. Cambridge Philos. Soc.* **100** (1986), 383–486.

[Tót01] Tóth, B., No more than three favorite sites for simple random walk. *Ann. Probab.* **29** (2) (2001), 484–503.

[Wer04] Werner, W., Random planar curves and Schramm-Loewner evolutions. In *École d'été de probabilitiés de St. Flour XXXII – 2002*, Lecture Notes in Math. 1840, Springer-Verlag, Berlin 2004, 113–196.

[Wil89] Wilf, H. S., The editor's corner: the white screen problem. *Amer. Math. Monthly* **96** (1989), 704–707.

Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.
E-mail: amir@math.stanford.edu

# Modelling genes: mathematical and statistical challenges in genomics

Peter Donnelly

**Abstract.** The completion of the human and other genome projects, and the ongoing development of high-throughput experimental methods for measuring genetic variation, have dramatically changed the scale of information available and the nature of the questions which can now be asked in modern biomedical genetics. Although there is a long history of mathematical modelling in genetics, these developments offer exciting new opportunities and challenges for the mathematical sciences. We focus here on the challenges within human population genetics, in which data document molecular genetic variation between different people. The explosion of data on human variation allows us to study aspects of the underlying evolutionary processes and the molecular mechanisms behind them; the patterns of genetic variation in different geographical regions and the ancestral histories of human populations; and the genetic basis of common human diseases. In each case, sophisticated mathematical, statistical, and computational tools are needed to unravel much of the information in the data, with many of the best methods combining complex stochastic modelling and modern computationally-intensive statistical methods. But the rewards are great: key pieces of scientific knowledge simply would not have been available by other means.

**Mathematics Subject Classification (2000).** Primary 92D10, 92D15; Secondary 65C05.

**Keywords.** Genetics, problems related to evolution, Monte Carlo methods.

## 1. Introduction

We begin with a brief review of the basic concepts and terminology from genetics. The full picture is both more complicated and richer than we need, and we present only a very high-level overview.

Genetic information is transmitted from parents to offspring, and carried in the nucleus of each cell, in DNA (deoxyribonucleic acid). To a mathematician, DNA can be thought of as a very long word over the four letter alphabet $\{A, C, G, T\}$, with each letter representing one of the four chemical bases, or nucleotides, which are arranged effectively linearly along the DNA molecule. It is the order in which the bases appear which conveys the information. Some parts of the molecule are "read" by molecular machinery, and the relevant part of the DNA is used as a template to make a particular protein. These parts of the DNA are called genes. The totality of an organism's DNA is called its genome. The human genome consists of about $3 \times 10^9$ bases, and contains around 25,000 genes, but most of the DNA in the human genome appears to have no

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

function. The genomes of different organisms differ in size, with some much smaller and some substantially larger than the human genome. Like many other organisms, humans are diploid, in that we carry two copies of our genome, one inherited from our mother, and one from our father. Human DNA is packaged into 23 pairs of chromosomes, with one copy of each chromosome inherited from each parent.

Each human sperm or egg (collectively referred to as germ cells) contains a single copy of each of the 23 chromosomes. For what follows, we need to understand a little about the process during which germ cells are formed, called meiosis. Focus on a particular human chromosome. The individual (progenitor) producing the germ cell will have two (slightly different) copies of this chromosome. Think of the process which produces the chromosome for the germ cell as starting on one of the chromosomes in the progenitor and copying from it along the chromosome. Occasionally, and for our purposes randomly, the copying process will "cross over" to the other chromosome in the progenitor, and then copy from that, perhaps later jumping back and copying from the original chromosome, and so on. The chromosome in the germ cell will thus be made up as a mosaic of the two chromosomes in the progenitor. The crossings over are referred to as recombination events. In a typical human meiosis, there will be only a few recombination events per chromosome. In addition to the process of recombination, there will be very occasional mutations: positions where the nucleotide in the offspring is different from that in the progenitor chromosome from which it is being copied. To give an idea of the scale of these effects in humans, the probability of a mutation in any particular nucleotide position is of order $10^{-8}$ per meiosis, and the average probability of a recombination event in a particular position is of the same order. Mutation and recombination are two of the fundamental evolutionary forces. Mutation introduces new variants into a population. (Some of these will make the resulting chromosome better at doing its job than the progenitor chromosome). The effect of recombination is more subtle, but equally important. Recombination allows the shuffling of variants between different backgrounds: when a mutation arises it occurs on a particular chromosome with a particular DNA sequence. Over generations, recombination events near this mutation allow it to be swapped onto different backgrounds.

In effect, the human genome project read one copy of the human genome ([12], [26]) – actually a mosaic made up of the genome from many individuals. The human genome sequence is available on the web, along with annotations which show, for example, which parts correspond to known and predicted genes, or regions which regulate the expression of genes, or appear to be highly conserved across species. The genomes of many other organisms are also now available, with more being completed each month. In each case a major challenge within the science is to better understand the function of, and interactions between, different parts of each of these genomes.

We can think of the human genome project as focussing on the aspects of our genome which we all share: the things that make us human. But there are also differences between people, in appearance, nature, abilities, and susceptibility to different diseases. Some of these differences have a genetic component, resulting

from differences in the DNA sequence between individuals. If we compared two human chromosomes in the same region then they would differ at about 1 place in 1000. (As a comparison, the human genome sequence differs from that of the chimpanzee at about 1 position in 100.)

Following on from the human genome project, there was a major effort in a public-private partnership to discover many of the positions at which human chromosomes differ. While there are a number of interesting ways in which DNA sequences can differ, the most common is when at a particular position, or nucleotide, some chromosomes in the population carry one letter (or base) while others carry a different letter (base). Such positions are called single nucleotide polymorphisms, or SNPs (pronounced "snips"). These SNPs are catalogued in public databases (e.g. http://www.bioinfo.org.cn/relative/dbSNP%20Home%20Page.htm). Over the current decade the number of SNPs known in humans has grown from hundreds to more than 8 million. Nonetheless, many remain undiscovered. For example it is estimated that there are 10 million "common" SNPs, that is SNPs where the rarer variant has a population frequency of at least 5% [11]. As noted above, the mutation rate at any particular nucleotide position is very small ($10^{-8}$). On the other hand the human genome is large ($3 \times 10^9$ nucleotides). SNPs can be thought of as positions at which mutations happen to have occurred in the genome, where the chromosome carrying the mutation has spread through the population. It is extremely rare for there to be more than two variants present at a particular SNP.

Having read one copy of the human genome sequence, then catalogued many of the DNA sequence variations present in human populations, a natural next step was to understand the patterns in which these variants occur in different populations. This has recently been undertaken by the International HapMap Consortium, a collaboration involving five different countries (Canada, China, Japan, UK, USA), at a cost of about $100M. Largely completed, the project typed around 3.5M SNPs in samples from four populations around the world: 90 Caucasians from Utah, 90 Yorubans from Ibadan, Nigeria, 45 Han Chinese from Beijing, and 45 Japanese from Tokyo. The first phase of the project, involving just over 1M SNPs, was reported in [11].

It turns out that the variants present at SNPs close to each other on the same chromosome are often correlated. That is, if at one SNP some chromosomes in the population carry an $A$ and others a $G$, while at a nearby SNP the two variants are $T$ and $C$, it might be that chromosomes which carry an $A$ at the first SNP are more likely to carry a $T$ at the second SNP than those with a $G$ at the first SNP. This kind of correlation is known in population genetics as linkage disequilibrium (LD). The correlations can be very strong (for example the extreme case where all chromosomes carry either $A$ and $T$ or $G$ and $C$, is not unusual) and as we will see below, are very important for studies of the genetics of human disease. Amongst other things, the HapMap project characterized patterns of linkage disequilibrium in the samples it studied.

The reasons for linkage disequilibrium are apparent when one thinks about the history of novel mutations. In the example above, suppose the mutation giving rise

to the first SNP ($A/G$) occurred further into the past than that giving rise to the second SNP, and suppose that at the second SNP the $C$ variant (variants are often called alleles in genetics) was the one present originally. The mutation creating a $T$ and giving rise to the second SNP will have occurred on a single chromosome in a particular generation. Suppose it occurred on a chromosome carrying an $A$ at the first SNP. Then it is immediate that when it arose, a $T$ at the second SNP would occur with an $A$ at the first SNP. Over subsequent generations, the number of copies of the chromosome carrying the $T$ is likely to have grown (otherwise it would not be present today) and unless there is a recombination event between the two SNP positions on one of these chromosomes, it will remain the case that a $T$ at the second SNP would always occur with an $A$ at the first SNP. This association will only be broken down by recombination events, and the extent of this will depend on two things: (i) how many nucleotides separate the two SNPs on the chromosome (the closer together, on average, the smaller is the chance of a recombination between them); and (ii) the number of generations since the mutation giving rise to the second SNP (since a larger number of generations will allow a greater chance for a recombination event). In general, the observed patterns of LD depend on a number of factors, including chance past recombination events, and the demographic history of the population concerned.

## 2. Mathematical models

The arguments in the previous section were entirely qualitative. While helpful, they do not allow quantitative assessments of the way in which various aspects of genetic variation depend on the underlying evolutionary forces or demographic effects. To do so requires the development and analysis of mathematical models of the evolutionary process.

There has been a long history of mathematical modelling in population genetics, dating back to Fisher and Wright early last century. For most questions of interest, stochastic effects are important and the principal models are probabilistic. For most of the period over which these models have been studied, empirical data against which to compare the models have been sparse. Typically, what data there were came from so-called model organisms (particular species of flies and worms for example). Over the last few years, there has been an explosion of data documenting genetic variation in humans, to the extent that our own species provides the richest setting in which to apply these models.

We aim here only to give a brief flavour of the stochastic models which arise in population genetics. The most basic models are finite Markov chains which describe the way in which the genetic composition of the population changes over time. In most cases, these models are not tractable, and interest moves to their limiting behaviour as the population size grows large, under suitable re-scalings of time. When examined forward in time, this leads to a nice family of measure-valued diffusions, called

Fleming–Viot processes. In a complementary, and for many purposes more powerful approach, one can instead look backwards in time, and focus on the genealogical tree relating sampled chromosomes. In the large population limit, these (random) trees converge to a particular process called the coalescent.

One simple discrete model for population demography is the Wright–Fisher model. Consider a population of fixed size $N$ chromosomes which evolves in discrete generations. (For many purposes it turns out that we can ignore the fact that chromosomes occur in pairs in individuals, and we do so here.) The random mechanism for forming the next generation is as follows: each chromosome in the next generation chooses a chromosome in the current generation (uniformly at random) and copies it, with the choices made by different chromosomes being independent. An equivalent description is that each chromosome in the current generation gives rise to a random number of copies in the next generation, with the joint distribution of these "offspring numbers" being symmetric multinomial. Under an assumption of genetic neutrality, all variants in a population are equally fit. In this case, one can first generate the demography of the population using, say, the Wright–Fisher model, and then independently superimpose the genetic type for each chromosome, and the details of the (stochastic) mutation process which may change types. The extent to which this neutrality assumption applies is rather controversial in general, and for humans in particular, but it seems likely that it provides a reasonable description for many parts of the genome. Recombination (and if needed natural selection) can be naturally added to the model. Where a recombination event occurs, the offspring chromosome will be made up from two chromosomes in the current population. Although we have described it in terms of chromosomes, it is natural only to apply the Wright–Fisher model to small regions of a chromosome. In this case, the probabilities of mutation and recombination in a copying event are both very small, and these events are rare.

The Wright–Fisher model may also be extended to allow for more realistic demographic effects, including variation in population size, and geographical spatial structure in the population (so that offspring chromosomes are more likely to be located near to their parents). We will not describe these here. Somewhat surprisingly, it transpires that the simple model described above, (constant population size, random mating, and neutrality – the so-called "standard neutral" model), or rather its large population limit, captures many of the important features of human evolution. There is an aphorism in statistics that "all models are false, but some are useful". The standard neutral model has proved to be extremely useful.

In a Wright–Fisher or any other model, we could describe the genetic composition of the population at any point in time by giving a list of the genetic types currently present, and the proportion of the population currently of each type. Such a description corresponds to giving a probability measure on the set $E$ of possible types. It is sometimes helpful to think of this measure as the distribution of the type of an individual picked at random from the population. In this framework, when we add details of the mutation process and recombination to the Wright–Fisher model, we obtain a discrete time (probability) measure-valued Markov process. As $N$ becomes

large a suitable rescaling of the process converges to a diffusion limit: time is measured in units of $N$ generations, and mutation and recombination probabilities are scaled as $N^{-1}$. For general genetic systems, the limit is naturally formulated as a measure-valued process, called the Fleming–Viot diffusion. The classical so-called Wright–Fisher diffusion is a one dimensional diffusion on $[0, 1]$ which arises when there are only two genetic types and one tracks the population frequency of one of the types. This is a special case of the Fleming–Viot diffusion, in which we can identify the value of the classical diffusion, $p \in [0, 1]$ with a probability measure on a set with just two elements. The beauty of the more general, measure-valued, formulation is that it allows much more complicated genetic types, which could track DNA sequences, or more exotically even keep track of the time since particular mutations arose in the population.

The Fleming–Viot process can thus be thought of as an approximation to a large population evolving according to the Wright–Fisher model. For the Wright–Fisher model, time is measured in units of $N$ generations in this approximation (and the approximation applies when mutation and recombination probabilities are of order $N^{-1}$). In fact the Fleming–Viot process arises as the limit of a wide range of demographic models, (and we refer to such models as being within the domain of attraction of the Fleming–Viot process) although the appropriate time scaling can differ between models. (See, for example, [5].) For background, including explicit formulations of the claims made above, see for example [2], [3] [4], [5], [6]. Donnelly and Kurtz ([2], [3]) give a discrete construction of the Fleming–Viot process. As a consequence, the process can actually be thought of as describing the evolution of a hypothetically infinite population, and it explicitly includes the demography of that population.

There has been considerable recent interest in looking backwards in time to study the genealogy of population genetics models. This is simplest in the absence of recombination. Consider again the discrete Wright–Fisher model. If we consider two different chromosomes in the current generation, they will share an ancestor in the previous generation with probability $1/N$. If not, they retain distinct ancestries, and will share an ancestor in the previous generation with probability $1/N$. The number of generations until they share an ancestor is thus geometrically distributed with success probability $1/N$ and mean $N$. In the limit for large $N$, with time measured in units of $N$ generations, this geometric random variable will converge to an exponential random variable with mean 1.

More generally, if we consider $k$ chromosomes, then for fixed $k$ and large $N$, they will descend from $k$ distinct ancestors in the previous generation with probability

$$1 - \binom{k}{2}\frac{1}{N} + O(N^{-2}).$$

Exactly two will share a common ancestor in the previous generation with probability $\binom{k}{2}\frac{1}{N} + O(N^{-2})$ and more than a single pair will share a common ancestor with probability $O(N^{-2})$. In the limit as $N \to \infty$, with time measured in units of $N$ generations, the time until any of the $k$ share an ancestor will be exponentially distributed

with mean $\binom{k}{2}^{-1}$, after which time a randomly chosen pair of chromosomes will share an ancestor.

Thus, in the large population limit, with time measured in units of $N$ generations, the genealogical history of a sample of size $n$, may be described by a random binary tree. The tree initially has $n$ branches, for a period of time $T_n$, after which a pair of branches (chosen uniformly at random independently of all other events) will join, or coalesce. More generally, the times $T_k$, $k = n, n-1, \ldots, 2$ for which the tree has $k$ branches are independent exponential random variables with

$$\mathrm{E}(T_k) = \binom{k}{2}^{-1},$$

after which a pair of branches (chosen uniformly at random independently of all other events) will join, or coalesce. The resulting random tree is called the $n$-coalescent, or often just the coalescent.

In a natural sense the tree describes the important part of the genealogical history of the sample, in terms of their genetic composition. It captures their shared ancestry, due to the demographic process. A key observation is that in neutral models the distribution of this ancestry is independent of the genetic types which happen to be carried by the individuals in the population. Probabilistically, one can thus sample the coalescent tree and then superimpose genetic types: first choose a type for the most recent common ancestor of the population (the type at the root of the coalescent tree) according to the stationary distribution of the mutation process, and then track types forward through the tree from the common ancestor, where they will possibly be changed by mutation.

The preceding recipe gives a simple means of simulating the genetic types of a sample of size $n$ from the population. Note that this is an early example of what has recently come to be termed "exact simulation": a finite amount of simulation producing a sample with the exact distribution given by the stationary distribution of a Markov process. In addition, it is much more computationally efficient than simulating the entire population forward in time for a long period and then taking a sample from it. Finally, it reveals the complex structure of the distribution of genetics models at stationarity – the types of each of the sampled chromosomes are (positively) correlated, exactly because of their shared ancestral history.

We motivated the coalescent from the Wright–Fisher model, but the same limiting genealogical tree arises for any of the large class of demographic models in the domain of attraction of the Fleming–Viot diffusion. Moreover, the way in which the tree shape changes under different demographic scenarios (e.g. changes in population size, geographical population structure) is well understood.

The discrete construction of the Fleming–Viot process described above actually embeds the coalescent and the forward diffusion in the same framework, so that one can think of the coalescent as describing the genealogy of a sample from the diffusion.

There is even a natural limit, as $n \to \infty$ of the $n$-coalescents. This can be thought of as the limit of the genealogy of the whole population, or as the genealogy of

the infinite population described by the Fleming–Viot process, although the analysis underlying the relevant limiting results is much more technical than that outlined above for the fixed-sample-size case. It is easiest to describe this tree from the root, representing the common ancestor of the population, forward to the tips, each of which represents an individual alive at the reference time. The tree has $k$ branches for a random period of time $T_k$, after which a branch, chosen uniformly at random, independently for each $k$, splits to form two branches. The times $T_k$, $k = 2, 3, \ldots$, are independent exponential random variables, and independent of the topology of the tree, with

$$\mathrm{E}(T_k) = \binom{k}{2}^{-1}.$$

Write

$$T = \sum_{k=1}^{\infty} T_k$$

for the total depth of the tree, or equivalently for the time back until the population first has a common ancestor. Note that $T$ is a.s. finite. In fact $\mathrm{E}(T) = 2$.

Now we return to the case where recombination is allowed. The simplest way to conceptualise this more general situation is that there is a genealogical tree, marginally distributed as the coalescent, associated with each nucleotide position. As one moves along the DNA sequence, these trees for different positions are highly positively correlated. In fact, two neighbouring positions will have the same tree iff there is no recombination event between those positions since their joint most recent common ancestor, on a lineage leading to the current sample. If there is such a recombination, the trees for the two positions will be identical back to that point, but (in general) different before it. The correlation structure between the trees for different positions is complex, and for example regarded as a process on trees as one moves along the sequence, it is not Markov. But it is straightforward to simulate from the relevant joint distribution of trees, and hence of sampled sequences. The trees for each position can be embedded in a more general probabilistic object (this time a graph rather than a tree) called the ancestral recombination graph ([8], [9]).

## 3. Disease mapping

One major current analytical challenge in the field is the development of statistical methods in genetic studies of human disease. A common study design is case-control: a (typically large) set of individuals with a particular disease (cases) and a set of healthy individuals (controls) are typed at a (large) set of SNPs. If one variant at a particular SNP predisposes individuals to (respectively protects them against) the disease in question then the frequency of that variant should be higher (lower) in the cases than the controls. The signal one looks for then is a difference in allele frequency between cases and controls at a particular SNP.

As one contemporary example, the Wellcome Trust Case Control Consortium is a large UK-based study in which 2000 cases for each of 8 common diseases (Type 1 and Type 2 Diabetes, Hypertension, Coronary Heart Disease, Crohn's Disease, Bipolar Disorder, Rheumatoid Arthritis, and Tuberculosis) will be compared with 3000 controls at around 500,000 SNPs. This size of study is becoming more common, and although expensive, is within reach of major biomedical research budgets. (For example, the study just described will cost around US$15M.)

Some human disorders have a simple genetic component. In these, a single gene will be involved, and mutations in that gene cause individuals to be affected. In some cases, such as Huntington's disease, individuals will be affected if either of their chromosomes carries the mutation. (The inheritance is said to be dominant.) In others, such as Cystic Fibrosis, individuals will be affected only if both their chromosomes carry mutations at the gene in question. (The inheritance is said to be recessive.) In these cases there is effectively a deterministic relationship between carrying mutated copies of the gene in question and having the disease. These so-called simple genetic diseases are typically rare, and often very debilitating. In a large number of cases the exact genes involved are now known.

Most or all of the common human diseases also have a genetic component, but one which acts in a more subtle, and complicated, way. We are some way from understanding the full story, but for these common human diseases, it is thought that mutations in genes may slightly increase (or decrease) the probability of the individual developing the disease, rather than deterministically predicting it. Disease susceptibility may well also involve the interaction between mutations in different genes, and/or interactions between genes and environmental or lifestyle factors.

One major issue with case-control studies involves the need for cases and controls to be as similar as possible apart from their disease status. A particular, genetic, concern relates to geographical population structure. Most human populations differ genetically – individuals from nearby geographical locations are more likely to be genetically similar than those from distant locations. This is well documented in comparisons between the major continental regions of the world. But the same effect pertains, to a lesser extent, within continental or even national regions. Suppose for simplicity that a population is actually made up of two subpopulations which differ genetically at a particular genetic marker (say the $A$ allele is more common in subpopulation 1 than in subpopulation 2), and that in addition, perhaps for environmental reasons, the disease is more common in subpopulation 1. Then a random sample of cases from the population as a whole will tend to include more individuals from subpopulation 1 than will a random sample of controls, and in turn the sample of cases will have a higher frequency of $A$ than will the controls. A naive analysis, which ignores the population substructure, might wrongly conclude that the $A$ variant at this SNP played a role in disease susceptibility.

This tendency for geographical population structure to lead to false positives in association studies actually led to the case-control design being largely ignored for many years. More recently a range of statistical approaches has been developed to

correct the problem. One class of approach uses all the markers typed to correct the null distribution of the usual test statistics. Another uses the markers to infer the underlying structure and assign individuals to subpopulations, with the comparisons between cases and controls being made only within subpopulations. Perhaps counter-intuitively, it is also the case that the problems caused by substructure increase with the size of the study: even the small amounts of structure within national populations might cause problems for the large studies currently being undertaken. See [19] for further discussion and additional references.

If an association study directly tests a SNP causatively involved in disease susceptibility, then we would expect to see frequency differences between cases and controls at that SNP. For common diseases the effect of carrying one variant is typically small, which will lead to only a small difference in the frequency of that variant between cases and controls. The large sample sizes of current studies are needed to ensure statistical power to detect small frequency differences.

Even were an investigator to restrict attention only to variants which occur at appreciable frequency (e.g. so-called "common variants", where the less common, or "minor" allele has frequency $> 5\%$), these cannot all be tested in an association study. It is estimated that there are probably around 10 million such variants in the human genome [11]. Firstly, many of these variants are not known, and secondly, the cost of checking all known variants is prohibitive (even at levels of current biomedical research funding).

Here, the correlations between alleles described above (recall the discussion of linkage disequilibrium) is very helpful. In an extreme case, suppose variants at two SNPs are perfectly correlated: each chromosome carrying an $A$ at the first SNP carries a $T$ at the second, and chromosomes with a $G$ at the first SNP carry a $C$ at the second. In this case, when an association study types one of the SNPs it effectively also types the other. Put another way, if one of the two SNPs were causatively involved in the disease, and a study typed the other SNP, then there should still be a signal.

In fact, this "extreme" case, of perfect correlation, is not uncommon. For example, it is estimated that in a Caucasian population, 60% of common SNPs have the property that there are at least three other SNPs with which they are perfectly correlated, and 20% are perfectly correlated with more than 20 other SNPs; only around 20% are not perfectly correlated with any other SNPs [11]. (There is in general less correlation between SNPs in samples from African populations.) The reasons for this are now well understood. They follow from properties of coalescent trees, and recently discovered facts about the human recombination process. Two SNPs will be perfectly correlated iff they occur on the same branch of the coalescent tree. The branches near the root of the tree are relatively long, thus allowing time for a number of mutations to occur. In addition, as we discuss in more detail below, it turns out that in humans, recombination events do not occur uniformly along the chromosome, but instead cluster into small regions, called recombination hotspots, which are typically widely separated. Between these regions there will often be effectively a single coalescent tree for all nucleotide positions, and the placement of mutations

on branches of this tree induces the correlations between SNPs.

Often two SNPs may be well correlated but not perfectly so. In this case, if one is causative and the other is typed, it may still be possible to see a signal, and hence to detect the untyped causative SNP. Under a simple disease model, this effect depends simply on the correlation coefficient, $r^2$, between the SNPs. If a sample size of $n$ were needed at given power to detect the SNP in a study in which it is typed directly, then a sample size of $n/r^2$ will be needed for the same power if only the correlated SNP were typed.

The major point of the HapMap project was to describe these correlations between SNPs in human populations. As a consequence, association studies can carefully choose which SNPs to type so as to minimize the number of SNPs involved. For example, it is estimated that in a Caucasian sample, genotyping a set of 300,000 SNPs will capture around 80% of all common SNPs with $r^2 > 0.8$ [11].

Whatever strategy is chosen to select SNPs for typing in an association study, the analytical challenge is how best to analyse the data. This can helpfully be thought of as a statistical missing data problem. We have data at a set of SNPs in cases and controls. For these SNPs we can directly test the possibility that their variants are related to disease susceptibility in a variety of ways (e.g. by simple chi-squared tests for differences in genotype frequencies, or by fitting say logistic regression models relating genotype to disease status). If we had data at the SNPs not typed in the study we could apply the same tests at those SNPs. Thinking of the data at the untyped SNPs as missing data, the challenge then is to learn about the missing data from the data we actually have. Some of the most promising approaches to this problem make use of the mathematical models described above, (or approximations to them). Informally, the models can be used to predict, or impute, data at the untyped SNPs from the data at the SNPs actually typed, and then this imputed data is used to test for a disease effect. The approach can be applied either at the positions of known SNPs not included in the study, or more generally at arbitrary positions in the genome. It combines the empirical information available from surveys such as HapMap, inferences as to recombination rates in the human genome, and population genetics models.

## 4. Human recombination

Recall that recombination is the process by which germ cells are constructed to contain part of each of the chromosomes in their progenitor. It has long been known that recombination events do not happen uniformly along the human chromosomes. (The rates of recombination even differ between males and females.) Over large scales, this can be seen in pedigree (or family) studies: effectively through localising the genomic positions of recombination events in comparisons between parents and their children.

But pedigree studies have limited resolution for estimating recombination rates. The average recombination rate across 10 million basepairs – 10 megabases (Mb) –

is $10^{-1}$. Reliable estimation of probabilities of order $10^{-1}$ requires many tens, or hundreds of observations. While this is realistic in human pedigree studies, the average recombination rate across 1Mb is an order of magnitude lower, and requires an order of magnitude more observations for accurate estimation, taking it effectively beyond the limit of practicability for pedigree studies. As a consequence, our understanding of the variation in human recombination rates based on pedigree studies does not go below the megabase scale.

So what do recombination rates look like over finer scales. Two recent lines of evidence suggested that the picture may be surprising and very interesting. The first was the direct observation of recombination hotspots: small (typically 2 kilobase, or kb) regions in which recombination events cluster, and for which the local recombination rate is much higher than in the surrounding DNA. These observations typically involved studies of human sperm. Although realistic pedigree studies are uninformative over these scales, clever and careful experimentation does allow detection of sperm with recombination events in particular small regions, from which recombination rates (in males) can be estimated over the region studied.

The second clue came from patterns of linkage disequilibrium (LD) in human populations. Contrary to the predictions of simple models, human linkage disequilibrium extended over much larger regions than expected, and in addition, the patterns showed regions of extended LD interrupted by short regions of LD breakdown. Although both the initial observations and the suggested causes were controversial, this pattern is now well documented in humans [11]. One natural explanation was that recombination events were not uniformly distributed but instead clustered into hotspots.

As we saw above, the patterns of genetic variation in human populations have been shaped by a number of effects, including recombination. In principle then, such data contain information about the underlying recombination rates. Armed with an understanding of the stochastic models of section 2, we could then treat this as a statistical problem, and try to use data to infer some of the parameters of the models, in particular the recombination rates.

It turns out that this is a challenging inference problem, for a number of reasons. In either classical or Bayesian statistical inference, a central role is played by the likelihood: the probability of the observed data as a function of model parameters. Although the stochastic models are well understood, and for example easy to simulate from, no explicit expressions are available for probabilities of interest, such as the stationary distribution. In the statistical context, this means that the likelihood associated with the model is not available analytically. One way of conceptualising the difficulty is as follows: for a given genealogical tree (or graph in the context of recombination) one could calculate the likelihood for given parameter values. The actual likelihood could then be obtained by averaging this quantity over all possible underlying trees or graphs. Herein lies the problem: the space of trees/graphs is so large that this averaging is impracticable.

Various clever computational approaches have been developed in modern statistics to overcome this type of problem, and there has been particular attention to these

in the population genetics context (see for example [25]). A general observation is that sophisticated understanding of the stochastic models allows big improvements in the quality and efficiency of the statistical inference. For example, for inference of recombination rates assumed constant across the region of interest, the best available methods are more efficient then their predecessors by up to four orders of magnitude [7].

But even the best available statistical methods based on the coalescent are impracticable for estimating recombination rates of interest for a different reason: the sheer size of available data sets. The Phase I HapMap data, for example, documents genetic variation at 1M SNPs in 269 individuals. Two different approaches have recently been developed to address this issue. In essence, each takes the view that the coalescent model is itself an approximation to reality, so why not make further judicious approximations in order to achieve tractability. One approach, pioneered by Li and Stephens [18] involves an alternative model for genetic data with a hidden Markov structure. (See [1] for application to the estimation of fine-scale recombination rates.) We concentrate here on the other approach, introduced by Hudson [10] in the context of constant recombination rates, and developed by McVean and colleagues for variable recombination rates [20], [21].

This approach retains the original coalescent model, and uses its exact likelihood for data at a pair of SNPs, as a function of the recombination rate between them. But rather than using the correct joint distribution for a set of more than two SNPs, the approximation assumes each pair of SNPs to be independent. In this way, a so-called pairwise composite likelihood is constructed as the product of the exact coalescent likelihoods across all pairs of SNPs in the data. In the setting of variable recombination rate, the parameter of interest is a piecewise constant function specifying the recombination rate between each pair of SNPs (so the function only changes value at the positions of SNPs). McVean *et al.* [21] adopt a Bayesian approach to inference. The prior distribution (here on function space) encourages smoothness in the rate function and reversible jump Markov chain Monte Carlo, using the pairwise composite likelihood, is used to explore the posterior distribution on recombination rates. One attractive feature of the prior distribution putting weight on smooth functions is that the approach "borrows" information from nearby SNPs – the estimated rate between a pair of SNPs will be influenced by data at nearby SNPs. Another positive consequence is the tendency to avoid overfitting: maximum likelihood estimation with the same likelihood function would fit a different recombination rate between each pair of SNPs. The paper also develops a formal likelihood ratio test for the presence of a recombination hotspot, based on the composite likelihood (with significance levels determined by appropriate simulation). Tests on real and simulated data show these methods to perform remarkably well. In spite of the approximations (firstly to give the coalescent and secondly of the coalescent) involved, the models appear to be capturing key features of the real world remarkably well.

These new statistical methods, applied to recent genome-wide variation data sets such as the HapMap, have enormously extended our knowledge of human recombi-

nation. For the first time, fine-scale estimates of recombination rates are available across the human genome. These so-called genetic maps have resolution 2-3 orders of magnitude finer than their predecessors from pedigree studies. They show striking variation in rates, by up to four orders of magnitude, over kilobase scales, and provide a powerful tool in studies of human disease. We now know that recombination hotspots are a ubiquitous feature of the human genome: whereas around 15-20 hotspots had been previously characterised, around 30,000 have been detected from the new approach, with an estimated average density of one hotspot per 50kb. The approach has demonstrated, also for the first time, that recombination hotspots are definitively a feature of female recombination, and more generally that the fine-scale recombination landscape appears similar between males and females. Contrary to previous reports, recombination rates are systematically lower within genes, but interestingly, systematically higher close to genes [11], [22].

One common general tool in genetics studies involves comparisons between species, and this has proved informative for recombination as well. Two comparisons between humans and our closest neighbouring species, the chimpanzee, revealed that recombination hotspots are a feature of chimpanzee recombination as well. But whereas human and chimpanzee DNA sequences agree at about 99% of nucleotide positions, the studies surprisingly found that the positions of recombination hotspots do not match between the two species, suggesting that they have evolved rapidly over evolutionary times [29], [23]. Comparisons between the time-averaged recombination rates estimated from population data and those in contemporary sperm suggest an even more rapid evolution of recombination hotspots, with substantial changes even over the half a million or so years over which human genetic variation has accumulated [13].

Perhaps the best example of mathematical and statistical approaches adding substantially to scientific knowledge in this area comes from studies of motifs associated with recombination hotspots. The question of why some parts of the DNA sequence act as recombination hotspots and some do not has been a major focus of research attention, and remains little understood. No clear pattern was available from the 15 hotspots directly characterised from human sperm typing. But by first identifying, and then studying, 25,000 hotspots, Myers *et al.* [22] were able to identify several short DNA sequence motifs associated with hotspots. (In fact the analysis was not straightforward, essentially because most hotspots are localised only to within 5-10kb by statistical methods. The key was to focus only on those hotspots containing specific sequences of several hundred basepairs – many such so-called repetitive elements abound in the human genome – and to compare them with the same sequences outside hotspots.) Although also an exciting ongoing story, this approach was the first to identify sequence motifs (one important example is the collection $CCTCCCT$ of eight basepairs) associated with human hotspots, and amongst the first to be identified for any organism.

## 5. Conclusion

In addition to explaining some of the science, our aim has been to give a sense of the central role being played in modern genomics by mathematical modelling and statistical methods. The Human Genome Project provided the foundation for a new generation of genetics research. As we build on that foundation, in our understanding of basic biology, of the genetic basis for disease susceptibility, and in the use of this information to develop new therapies and preventions for human disease, it is clear that the mathematical and computational sciences will continue to play a vital role.

## References

[1] Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., Stephens, M., Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* **36** (2004), 700–706.

[2] Donnelly, P., Kurtz, T. G., A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Probab.* **24** (2) (1996), 698–742.

[3] Donnelly, P., Kurtz, T. G., Particle representations for measure-valued population models. *Ann. Probab.* **27** (1) (1999), 166–205.

[4] Ethier, S. N., Kurtz, T. G., *Markov processes. Characterization and convergence.* Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley, New York 1986.

[5] Ethier, S. N., Kurtz, T. G., Fleming-Viot processes in population genetics. *SIAM J. Control Optim.* **31** (2) (1993), 345–386.

[6] Ewens, W. J., *Mathematical population genetics. I. Theoretical introduction*. Second edition, Interdiscip. Appl. Math. 27, Springer-Verlag, New York 2004.

[7] Fearnhead, P., Donnelly, P., Estimating recombination rates from population genetic data. *Genetics* **159** (2001), 1299–1318.

[8] Griffiths, R. C. and Marjoram, P., Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3** (1996), 479–502.

[9] Griffiths, R. C. and Marjoram, P., An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (ed. by Peter Donnelly and Simon Tavaré), IMA Vol. Math. Appl. 87, Springer-Verlag, Berlin 1997, 257–270.

[10] Hudson, R. R., Two-locus sampling distributions and their application. *Genetics* **159** (2001), 1805–1817.

[11] The International HapMap Consortium, A haplotype map of the human genome. *Nature* **437** (2005), 1299–1320.

[12] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409** (2001), 860–921.

[13] Jeffreys A. J., Neumann R., Panayi M., Myers S., Donnelly P., Human recombination hot spots hidden within regions of strong marker association. *Nature Genetics* **37** (2005), 601–606.

[14] Kingman, J. F. C., Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics* (Rome, 1981), North-Holland, Amsterdam, New York 1982, 97–112.

[15] Kingman, J. F. C., The coalescent. *Stochastic Process. Appl.* **13** (3) (1982), 235–248.

[16] Kingman, J. F. C., On the genealogy of large populations. Essays in statistical science. *J. Appl. Probab.* Special Vol. **19A** (1982), 27–43.

[17] Kurtz, T. G., Martingale problems for conditional distributions of Markov processes. *Electron. J. Probab.* **3** (9) (1998), 29 pp. (electronic).

[18] Li, N., Stephens, M., Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165** (2003), 2213–2233.

[19] Marchini, J., Cardon, L., Phillips, M., Donnelly P., The effects of human population structure on large genetic association studies. *Nature Genetics* **36** (2004), 512–517.

[20] McVean, G. A. T., Awadalla, P., Fearnhead, P., A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160** (2002), 1231–1241.

[21] McVean, G. A. T., Myers, S., Hunt, S., Deloukas, P., Bentley, D. R., Donnelly, P., The fine-scale structure of recombination rate variation in the human genome. *Science* **304** (2004), 581–584.

[22] Myers S., Bottolo L., Freeman C., McVean G., Donnelly P., A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310** (2005), 321–324.

[23] Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A., Pääbo, S., Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics* **37** (2005), 429–434.

[24] Saunders, I. W., Tavaré, S., Watterson, G. A., On the genealogy of nested subsamples from a haploid population. *Adv. in Appl. Probab.* **16** (3) (1984), 471–491.

[25] Stephens, M., Donnelly, P., Inference in Molecular Population Genetics. *J. Royal Statist. Soc. Ser. B* **62** (2000), 605–655.

[26] Venter, J. C. *et al.*, The sequence of the human genome. *Science* **291** (2001), 1304–1354.

[27] Watterson, G. A., Mutant substitutions at linked nucleotide sites. *Adv. in Appl. Probab.* **14** (2) (1982), 206–224.

[28] Watterson, G. A., Substitution times for mutant nucleotides. Essays in statistical science. *J. Appl. Probab.* Special Vol. **19A** (1982), 59–70.

[29] Winckler W., Myers S. R., Richter D. J., Onofrio R. C., McDonald G. J., Bontrop R. E., McVean G. A. T., Gabriel S. B., Reich D., Donnelly P., Altshuler D., Fine-scale recombination rates differ markedly in human and chimpanzee. *Science* **308** (2005), 107–111.

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
E-mail: donnelly@stats.ox.ac.uk

# Geometric stochastic analysis on path spaces

K. David Elworthy and Xue-Mei Li*

**Abstract.** An approach to analysis on path spaces of Riemannian manifolds is described. The spaces are furnished with 'Brownian motion' measure which lies on continuous paths, though differentiation is restricted to directions given by tangent paths of finite energy. An introduction describes the background for paths on $\mathbb{R}^m$ and Malliavin calculus. For manifold valued paths the approach is to use 'Itô' maps of suitable stochastic differential equations as charts. 'Suitability' involves the connection determined by the stochastic differential equation. Some fundamental open problems concerning the calculus and the resulting 'Laplacian' are described. A theory for more general diffusion measures is also briefly indicated. The same method is applied as an approach to getting over the fundamental difficulty of defining exterior differentiation as a closed operator, with success for one and two forms leading to a Hodge–Kodaira operator and decomposition for such forms. Finally there is a brief description of some related results for loop spaces.

**Mathematics Subject Classification (2000).** Primary 58B10, 58J65; Secondary 58A14, 60H07, 60H10, 53C17, 58D20, 58B15.

**Keywords.** Path space, Hodge–Kodaira theory, infinite dimensions, connection, de Rham cohomology, stochastic differential equations, Malliavin calculus, Sobolev spaces, abstract Wiener spaces, differential forms.

## 1. Introduction

**1.1. Analysis with Gaussian measures.** Classical differential and geometric analysis is based on Lebesgue measure. The non-existence of an analogue of Lebesgue measure in infinite dimensions is demonstrated by the following theorem:

**Theorem 1.1.** *If $\mu$ is a locally finite Borel measure on a separable Banach space $E$ such that translations by every element of $E$ preserve sets of measure zero, then either $\mu = 0$ or $E$ is finite dimensional.*

'Local finiteness' here means that every point of $E$ has a neighbourhood with finite measure. The theorem is a special case of more general results, e.g. see Theorem 17.2 of [62]. In a sense it is behind many of the mathematical difficulties in 'path integration' and has meant that infinite dimensional differential and geometric analysis has had to develop its own techniques.

*Xue-Mei Li (-Hairer) has benefited from a Royal Society Leverhulme senior research fellowship.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The analysis proposed by Gross was based on his notion of *abstract Wiener spaces*. These are triples $\{i, H, E\}$ where $i\colon H \to E$ is a continuous linear injective map, with dense range, of a Hilbert space $(H, \langle\,,\,\rangle_H)$ into a separable Banach space $E$. (Throughout this article *all the linear spaces considered will be real*.) The defining property is that there is a Borel measure $\gamma$, say, on $E$ whose Fourier transform $\hat{\gamma}$ is given by

$$\hat{\gamma}(l) := \int_E e^{\sqrt{-1}l(x)}\, d\mu(x) = e^{-\frac{|j(l)|_H^2}{2}}$$

for all $l \in E^*$ where $j\colon E^* \to H$ is the adjoint of $i$. This was a generalisation of *classical Wiener space* where some analysis had been previously investigated, particularly by Cameron & Martin e.g. see [8], and also was influenced by Irving Segal's work. It was shown later that all, centred and strictly positive, so-called 'Gaussian measures' on a separable Banach space $E$ arise from an essentially unique abstract Wiener space structure on $E$, e.g. see [14].

Classical Wiener space can be considered as the special case when $E$ is the space $C_0([0, T]; \mathbb{R}^m)$ of continuous maps of a fixed interval $[0, T]$ into $\mathbb{R}^m$ which start at the origin, and $H$, sometimes called the *Cameron–Martin space*, is the space of finite energy paths $L_0^{2,1}([0, T]; \mathbb{R}^m)$, i.e. those paths in $C_0([0, T]; \mathbb{R}^m)$ which have distributional derivatives in $L^2$. The map $i$ is the inclusion. The norm for $H$ is given by $|h|_H^2 = \int_0^t |\dot{h}(t)|^2 dt$, and the measure on $E$ is the classical Wiener measure constructed by Wiener, so that the canonical process $[0, T] \times C_0([0, T]; \mathbb{R}^m) \to \mathbb{R}^m$ given by evaluation, is the standard model of Brownian motion. Denote that measure by $\boldsymbol{P}$.

The starting point for Gross's analysis was his extension of Cameron & Martin's quasi-invariance theorem to the case of abstract Wiener spaces and their measures $\gamma$:

**Theorem 1.2.** *Translation by an element $v$ of $E$ preserves sets of measure zero if and only if $v$ lies in the image of $H$. Moreover, for $h \in H$ and integrable $f\colon E \to \mathbb{R}$ we have for any $t \in \mathbb{R}$:*

$$\int_E f(x)\, d\gamma = \int_E f(x + ti(h)) \exp\Big( - t\mathcal{P}(h) - \frac{t^2}{2}|h|_H^2 \Big)\, d\gamma, \tag{1}$$

*where $\mathcal{P}\colon H \to L^2(E; \mathbb{R})$ is the* Paley–Wiener *map.*

The Paley–Wiener map is an isometry into $L^2$ defined as the $L^2$-limit of any sequence, $\{l_n\}_{n\geq 1}$ of elements in $E^*$ for which $\{j(l_n)\}_{n\geq 1}$ converges in $H$ to $h$. For classical Wiener space it is written as $\sigma \mapsto \int_0^T \langle \dot{h}(s), d\sigma(s)\rangle$ and called the *Paley–Wiener* (*stochastic*) *integral*. If $E = H = \mathbb{R}^n$ then $\mathcal{P}(h)(x) = \langle h, x\rangle_H$ which accounts for the notation $\langle h, x\rangle\tilde{}_H$ sometimes used for it in general.

The Gross–Cameron–Martin formula was given here with a parameter $t$ in order to obtain an integration by parts formula from it by differentiating with respect to $t$ at $t = 0$. If $f$ is sufficiently regular, for example Fréchet differentiable and bounded with bounded derivative $Df : E \to E^*$, this yields:

**Corollary** (Integration by parts). *For all* $h \in H$

$$\int_E Df(x)(i(h))\, d\gamma(x) = -\int_E f(x)\, \mathrm{div}(h)(x)\, d\gamma(x) \tag{2}$$

*where* $\mathrm{div}(h) = -\mathcal{P}(h)$.

Diffeomorphisms of the form $x \mapsto x + ij\alpha(x)$ between open subsets of $E$ where $\alpha$ is a $C^1$ map into $E^*$ were shown by H.-H. Kuo to also preserve sets of measure zero. This led him, starting in his thesis, and then others, to the study of *abstract Wiener manifolds*: Banach manifolds modelled on the space $E$ of an abstract Wiener space whose interchange of charts were of Kuo's form, [42]. Such manifolds have a natural class of Borel measures, locally equivalent to $\gamma$, and many of the usual constructions and results of the finite dimensional situation go over to them, [43], [20], [18], [55].

For any abstract Wiener space the map $i$ is compact. It follows that the derivatives of transformations of Kuo's type are linear maps which differ from the identity by a compact operator. This implies that abstract Wiener manifolds are Fredholm manifolds, [18]. For a wide class of Banach spaces $E$ the theory and classification of such manifold structures showed that every separable metrisable manifold $M$ modelled on $E$ can be given the structure of an abstract Wiener manifold, with the K-theory of $M$ playing the major role in their classification, [17].

It soon became clear that although interesting manifolds, such as path and loop spaces on finite dimensional manifolds admit these structures, in most interesting cases there is no natural one. Exceptions are finite codimensional submanifolds of abstract Wiener spaces, such as the space of paths from one submanifold embedded in $\mathbb{R}^m$ to another. Also see [32]. More general transformations preserving sets of measure zero were described, notably by Ramer in 1974, and then using Malliavin calculus by Kusuoka in 1982, and for flows of a class of vector fields on classical Wiener space by Driver, [12]. However the form of these transformations is not so different from those of Kuo, though the identity map in the decomposition may be replaced by a 'rotation'. See [61]. This together with the advent of Malliavin calculus, in 1976, with emphasis on mappings determined by stochastic differential equations, led to a move away from this approach, or at least a major modification of it [44], [45].

In [36] Gross shows that for any abstract Wiener space $\{i, H, E\}$ there is an abstract Wiener space $\{i', H, E'\}$ and a compact linear map $k\colon E' \to E$ such that $i = k \circ i'$. In other words, in the infinite dimensional case the measure $\gamma$ can be considered to lie in a smaller space than $E$; (however $H$ itself has measure zero). In the classical case this is demonstrated by the fact that the space of continuous functions can be replaced by the closure of $L_0^{2,1}$ in the space of Hölder continuous functions of exponent $\alpha$ for any $0 < \alpha < 1/2$. In Malliavin calculus on these linear spaces the space $E$ loses its importance, and in some treatments essentially disappears, e.g. see [39], and [50]. In the latter it is the Paley–Wiener functions, $\{\mathcal{P}(h) : h \in H\}$ which play the dominant role, returning to Segal's 'weak distribution' theory, [57]. However in the non-linear case of diffusion measures on path spaces of manifolds it seems necessary, at least at

the moment, to deal with the actual manifold on which the measures sit, though this could be taken to be Hölder continuous paths rather than continuous paths if that is more convenient. The treatment of Malliavin calculus below is organised with this in mind.

**1.2. Malliavin calculus on $E$.** From Gross's work, especially [35], it was clear that the basic differentiation operator on an abstract Wiener space should be the *$H$-derivative*. This could be defined on a basic domain, $\mathrm{Dom}(d_H)$, of functions $f\colon E \to \mathbb{R}$ consisting of a set of Fréchet differentiable functions which is dense in $L^p$ and whose $H$-derivatives: $d_H f\colon E \to H^*$ given by $d_H f_x(h) = Df(x)(i(h))$ lie in $L^p$, for all $1 \le p < \infty$. The integration by parts formula, equation (2), implies that $d_H$ is closable as a map between $L^p$ spaces with closure a closed linear map

$$d\colon \mathrm{Dom}(d) \subset L^p(E;\mathbb{R}) \to L^p(E;H^*).$$

Let $\mathbb{D}^{p,1}$ denote $\mathrm{Dom}(d)$ with its graph norm.

Our Paley–Wiener functionals, $\mathcal{P}(h)$, are easily seen to be in $\mathbb{D}^{p,1}$ for all $1 \le p < \infty$ with $d\mathcal{P}(h)_x(k) = \langle h, k\rangle_H$ for all $x \in E$ and $k \in H$, despite their lack of continuity in $E$. In fact the main point of the theory is that, for classical Wiener space, more general stochastic integrals and solution maps of stochastic differential equations, as described below, all lie in these Sobolev spaces.

The following characterisation of $\mathbb{D}^{p,1}$ for $1 < p < \infty$ was given by Sugita:

**Theorem 1.3** ([60]). *If $f \in L^p(E;\mathbb{R})$ then $f \in \mathbb{D}^{p,1}$ if and only if both of the following hold.*

1. *For each $h \in H$ there is a function $f_h\colon E \times \mathbb{R} \to \mathbb{R}$ which is absolutely continuous in the second variable and has $f_h(x,t) = f(x+th)$, for almost all $x \in E$, for each $t \in \mathbb{R}$.*
2. *There exists $df \in L^p(E;H^*)$ such that for any $h \in H$, $\frac{1}{t}(f(x+th) - f(x))$ converges in measure to $df_x(h)$ as $t \to 0$.*

From this we see that the spaces $\mathbb{D}^{p,1}$, for $1 < p < \infty$ are independent of any reasonable choice of initial domain $\mathrm{Dom}(d_H)$. A comforting fact; but one which is still open in the corresponding situation for paths on curved spaces, as will be described below.

For functions with values in a separable Hilbert space $G$ the spaces $\mathbb{D}^{p,1}(E;G)$ are defined in the analogous way, with the derivative $df$ now mapping $E$ into $\mathcal{L}_2(H;G)$, the space of Hilbert–Schmidt operators of $H$ into $G$. This is a Hilbert space often identified with the completed tensor product $G \otimes H$. It occurs because a basic property of an abstract Wiener space is that any continuous linear map from $E$ to a Hilbert space $G$, such as $Df(x)$ if $f\colon E \to G$ is Fréchet differentiable, gives a Hilbert–Schmidt operator when composed with $i$, e.g. see Thm 17.3 in [62]. We can iterate this procedure to obtain higher order Sobolev spaces.

As usual the gradient can be defined for functions in $\mathbb{D}^{p,1}$, $1 < p < \infty$, by $\langle \nabla f(x), h\rangle_H = df_x(h)$ to give an *H-vector field*, $\nabla f : E \to H$. It is a closed operator between the $L^p$ spaces with the negative of its adjoint denoted by div, a closed operator from Dom(div) in $L^q(E; H)$ to $L^q(E; \mathbb{R})$, where $q$ is the conjugate of $p$. Similar we have the adjoint $d^*$ of $d$. From this we get the analogue of the (Witten) Laplacian, or the 'Ornstein–Uhlenbeck operator', $\mathcal{L} = \operatorname{div} \nabla = -d^*d$. In the case $E = H = \mathbb{R}^n$ this is given by

$$\mathcal{L}(f)(x) = \Delta f(x) - Df(x)(x)$$

for $\Delta$ the usual Laplacian (with negative spectrum) of $\mathbb{R}^n$.

The Ornstein–Uhlenbeck operator acting on $L^2$ is the well known operator whose spectrum consists of 0 as unique ground state, together with the negative integers as eigenvalues of infinite multiplicity, corresponding to the homogeneous chaos decomposition of $L^2(E; \mathbb{R})$, and conjugate to the number operator of mathematical physics acting on the real symmetric Fock space. For example from above we see that for $h \in H$ the map $\mathcal{P}(h)$ is an eigenvector of eigenvalue minus one (so giving the 'one-particle' space). For more, see for example [37], [53], or [38].

For classical Wiener space an H-vector field $V : C_0 \to L_0^{2,1}$ is said to be *non-anticipating* if for each time t its value $V(\sigma)_t$ at the path $\sigma$ depends only on the restriction of $\sigma$ to the interval $[0, T]$. If this holds and it is in $L^2$, then it is in the domain of the divergence operator and $\operatorname{div}(V)(\sigma)$ is precisely the negative of the *Itô stochastic integral*, $\int_0^T \dot{V}(\sigma)_t \, d\sigma(t)$, as shown by Gaveau. This is the integral which is the basis of stochastic calculus. In the anticipating case it is the *Skorohod*, or *Ramer–Skorohod integral*, now by definition: although here the word 'integral' can be misleading since, as in finite dimensions, differentiation may be involved, [53].

An $L^2$-deRham and Hodge–Kodaira theory was given in this context by Shigekawa [58]. The $k$-forms were '$H$-forms', i.e. maps from $E$ into $\wedge^k H$ where $\wedge^k$ denotes the Hilbert space completion of the $k$-th exterior power, with the exterior derivative being a closed operator derived from our $H$-derivative $d$. The Hodge decomposition was just as in finite dimensional, standard, $L^2$-theory, and Shigekawa proved a vanishing theorem, implying the expected triviality of the deRham cohomology. A theory of finite co-dimensional forms was proposed by Ramer in his Thesis, in the context of abstract Wiener manifolds; further developments were made by Kusuoka, [46], but more is needed to develop the theory, even on domains in these linear spaces.

## 2. Scalar analysis on paths in *M*

**2.1. Brownian motion measure and Bismut tangent spaces.** Consider a smooth manifold $M$. For a fixed time $T > 0$, and a fixed $x_0 \in M$ let $C_{x_0}([0, T]; M)$, or simply $C_{x_0}$, denote the space of continuous paths $\sigma : [0, T] \to M$ starting at $x_0$, together with its usual $C^\infty$ Banach manifold structure, e.g. see [19] or [54]. The

tangent space $T_\sigma C_{x_0}$ to $C_{x_0}$ at a point $\sigma$ can be identified with the space of continuous paths $v\colon [0,T] \to TM$ into the tangent bundle to $M$, such that $v(0) = 0$ and $v(t) \in T_{\sigma(t)}M$ for $0 \le t \le T$. For a complete Riemannian manifold the *Brownian motion measure*, $\mu_{x_0}$, on $C_{x_0}$ is the unique Borel measure for which

$$\begin{aligned}\mu_{x_0}(\{\sigma \in C_{x_0} : \sigma(t_j) \in A_j,\, j = 1, 2, \dots, k\}) \\ = \int_{A_1}\int_{A_2}\cdots\int_{A_n} \prod_{j=0}^{j=k-1} p_{t_{j+1}-t_j}(x_j; dx_{j+1})\end{aligned} \tag{3}$$

where $0 = t_0 < t_1 < \cdots < t_k \le T$, the $A_j$ are Borel subsets of $M$, and the measures $p_t(x, dy)$ are the heat kernel measures: $p_t(x, dy) = p_t(x, y)dy$ for $p_t(x, y)$ the fundamental solution of the heat equation $\frac{\partial f}{\partial t} = \frac{1}{2}\Delta$ for $\Delta$ the Laplace Beltrami operator, div grad, of $M$.

For simplicity we shall assume that $M$ is compact. Let its dimension be $n$.

From the successes of the flat space case it was expected that, to do analysis on the path space $C_{x_0}$ using Brownian motion measure, the differentiation should only take place in a special set of directions. In the case of Gaussian measures on linear space a natural choice was given, as described above, by the linear structure together with the measure: but there are other choices as we see in Section 2.3 below and it is not clear if the measure plus the differential structure does determine a special one, cf. [21]. Nevertheless a natural choice for Brownian motion measure is the *Bismut tangent spaces*. These are Hilbert spaces, $\mathcal{H}_\sigma$, of tangent vectors, defined for almost all $\sigma \in C_{x_0}$ by

$$\mathcal{H}_\sigma = \{v \in T_\sigma C_{x_0} : (/\!/_\cdot)^{-1} v(\,\cdot\,) \in L^{2,1}([0,T]; T_{x_0}M)\} \tag{4}$$

where $/\!/_t$ denotes parallel translation along $\sigma$ using the Levi-Civita connection.

Because our paths $\sigma$ are typically so irregular, e.g. almost surely $\alpha$-Hölder continuous only for $\alpha < 1/2$, the parallel translation has to be constructed by stochastic differential equations and so is only defined along almost all paths. However if we set $\mathcal{H} = \cup_\sigma \mathcal{H}_\sigma \subset TC_{x_0}$ we will see that it has the rudiments of a bundle structure. We will call its sections $H$-vector fields, and the sections of its dual bundle $\mathcal{H}^*$ will be called $H$-one-forms, cf. [40].

In [12], Driver extended Cameron–Martin's theorem and the formulae (1), and (2) to this situation, showing that if $V^h$ is the $H$-vector field whose value at $\sigma$ is obtained by parallel translation of a fixed element $h \in L^{2,1}([0,T]; T_{x_0}M)$ along $\sigma$ then this measurable vector field has a solution flow which preserves sets of $\mu_{x_0}$-zero, with consequent analogues of equations (1) and (2).

As for flat space the integration by parts formula gives closability of the $H$-derivative $d_H\colon \mathrm{Dom}(d_H) \to L^2\Gamma\mathcal{H}^*$ from its domain in $L^2$ into the $L^2$-$H$-one-forms. It works for the $L^p$-spaces but we shall only mention $L^2$ from now on for simplicity. A natural, essentially the smallest natural, domain to choose is to let $\mathrm{Dom}(d_H)$ be the space $\mathrm{Cyl}(M)$ of smooth cylinder functions: those maps of the form

$\sigma \mapsto F(\sigma(t_i), \ldots, \sigma(t_k))$ for some smooth $F$ defined on the $k$-fold product of $M$, some $0 \leq t_1 < \cdots < t_k \leq T$, any natural number $k$. Other choices include the space of (Fréchet) $C^1$- functions which are bounded together with their derivatives, using the natural Finsler metric on $C_{x_0}$. However this time we do not know if these lead to the same domain for the closure of $d_H$, see [26], [27].

We must make a choice, and will choose $\mathrm{Cyl}(M)$ as basic domain. With this choice let $\mathbb{D}^{2,1}(C_{x_0})$, or $\mathbb{D}^{2,1}$, denote the domain of the $L^2$-closure of $d_H$ with its graph norm, with $\mathbb{D}^{2,1}(C_{x_0}; G)$ for the corresponding space of $G$-valued functions, $G$ a separable Hilbert space. Let $d$ denote the closure of $d_H$, so if $f : C_{x_0} \to G$ is in $\mathbb{D}^{2,1}(C_{x_0}; G)$ then $df$ is an $L^2$-section of $\mathcal{L}_2(\mathcal{H}; G)$ the 'bundle' with fibre at $\sigma$ the space of Hilbert–Schmidt maps of $\mathcal{H}_\sigma$ into $G$, sometimes denoted by $\mathcal{G} \otimes \mathcal{H}$.

With this we get a closed operator $\nabla$ as usual, mapping its domain $\mathbb{D}^{2,1}$ into $H$-vector fields, with adjoint the negative of a closed operator div. As usual we have a self adjoint 'Laplacian', or 'Ornstein–Uhlenbeck' operator $\mathcal{L}$ acting on functions, defined by $\mathcal{L} = \operatorname{div} \nabla = -d^* d$. The associated Dirichlet forms and processes have been studied, e.g. [13], [15]. Norris devised a stochastic partial differential equation to construct associated 'Brownian motions' or 'Ornstein–Uhlenbeck processes' on these path spaces, treating them as two parameter $M$-valued processes, [52]. The existence of a spectral gap for $\mathcal{L}$ was proved by S. Fang, and Log Sobolev inequalities independently by E. Hsu and Aida & Elworthy, see [37], [22]. However little, if anything, appears to be known else about its spectrum.

To discuss higher derivatives it is convenient to have a 'connection' on $\mathcal{H}$ in order to differentiate its sections. The most obvious choice is to use the trivialisation of $\mathcal{H}$ obtained simply by parallel translating every element in each $\mathcal{H}_\sigma$ back to an element of $L^{2,1}([0, T]; T_{x_0}M)$, so that $H$-vector fields can be considered as maps of $C_{x_0}$ into $L^{2,1}([0, T]; T_{x_0}M)$ to which we may try to apply our closed derivative operator $d$. This approach was used effectively, for example in [47]. However it does not conserve the $C_{\mathrm{Id}}([0, T]; \mathrm{GL}(n))$-structure of our path space, nor as Cruziero& Malliavin pointed out, does it fit well with the underlying 'Markovianity' of our set up. This led them to the 'Markovian' connection, see [11], a modification of which we will describe below.

**2.2. Itô maps and the stochastic development.** The stochastic development map $\mathcal{D} : C_0([0, T]; T_{x_0}M) \to C_{x_0}$ is an almost surely defined version of the Cartan development, describing 'rolling without slipping' along smooth paths. Its inverse is given by $\mathcal{D}^{-1}(\sigma)(t) = \int_0^t (/\!\!/_t)^{-1} \circ d\sigma(t)$, where the integral is a Stratonovitch stochastic integral, and $/\!\!/_t$ refers to parallel translation along the path $\sigma$, (defined for almost all paths). Reformulating Gangolli, [33], [18], it was shown by Eells & Elworthy that it sends Wiener measure to the Brownian motion measure. A fundamental result of Malliavin calculus is that, for each time $t$ the map can be $H$-differentiated infinitely often in the Sobolev sense. This was used by Driver to transfer his results about flows of vector fields, and integration by parts formulae, from flat space to $C_{x_0}$, see [12] where background details are included. However the use of $\mathcal{D}$ as a chart was lim-

ited because its $H$-derivative does not map $L_0^{2,1}([0,T];T_{x_0}M)$ to the Bismut tangent spaces. Furthermore from [49] it now seems that, unless $M$ is flat, composition with $\mathcal{D}$ will not pull elements in $\mathbb{D}^{2,1}(C_{x_0})$ back to elements in the domain of $d$: there will be a loss of differentiability.

An alternative technique is to use the solution maps, *Itô maps*, of more simple stochastic differential equations as replacements for charts. For this take a (Stratonovich) stochastic differential equation

$$dx_t = X(x_t) \circ dB_t + A(x_t)dt \tag{5}$$

on $M$. Here $A$ is a smooth vector field and $X$ gives linear maps $X(x)\colon \mathbb{R}^m \to T_xM$, smooth in $x \in M$. Also $B$ is the canonical Brownian motion given by $B_t\colon C_0([0,T];\mathbb{R}^m) \to \mathbb{R}^m$ with $B_t(\omega) = \omega(t)$ for $C_0([0,T];\mathbb{R}^m)$ furnished with its Wiener measure, which we shall now denote by $\boldsymbol{P}$.

The solution $x_t\colon C_0([0,T];\mathbb{R}^m) \to M$ to such an equation, starting from $x_0$, can be obtained by 'Wong–Zakai approximation': taking piecewise linear approximations $B_t^\Pi$ to the Brownian motion, for each partition $\Pi$ of $[0,T]$, and solving the family of ordinary differential equations

$$\frac{dx^\Pi(\omega)}{dt} = X(x_t^\Pi(\omega))\frac{dB^\Pi(\omega)}{dt} + A(x_t^\Pi(\omega))$$

starting at $x_0$, for each $\omega$. The required solution $x_t$ is given by $x_t(\omega) = \mathcal{I}(\omega)_t$ for $\mathcal{I}$ the limit in probability of $x^\Pi\colon C_0([0,T];\mathbb{R}^m) \to C_{x_0}$ as the mesh of $\Pi$ goes to zero. The map $\mathcal{I}$ is the *Itô map*. To be precise we have to choose it as a representative from an almost sure equivalence class of measurable maps. However, as with the stochastic development these maps can be differentiated arbitrarily many times in the sense of Malliavin calculus. In particular for almost all $\omega$ there is a linear $H$-derivative $T_\omega\mathcal{I}\colon H \to T_{\mathcal{I}(\omega)}C_{x_0}$.

The solutions to equation (5) form a Markov process with generator $\mathcal{A}$ where

$$\mathcal{A} = \frac{1}{2}\sum_{j=1}^{m} \mathcal{L}_{X^j}\mathcal{L}_{X^j} + \mathcal{L}_A. \tag{6}$$

For them to be Brownian motions we need $\mathcal{A} = \frac{1}{2}\Delta$ which requires each mapping $X(x)\colon \mathbb{R}^m \to T_xM$ to be surjective and induce the given Riemannian metric on the tangent space, or equivalently for the adjoint $Y_x\colon T_xM \to \mathbb{R}^m$ of $X(x)$ to be a right inverse of $X(x)$. Given that, we may choose the vector field $A$ appropriately. Then $\mathcal{I}$ will map the flat Wiener measure $\boldsymbol{P}$ to our Brownian motion measure $\mu_{x_0}$. In general the dimension, $m$, of the space on which the driving Brownian motion runs, will be larger than that of $M$ so that $\mathcal{I}$ will not be injective. The disadvantage of this can be reduced by 'filtering out the redundant noise' and to do this successfully we need to note that our SDE for Brownian motion determines a metric connection, $\breve{\nabla}$ on $TM$ by using $X$ to project the trivial connection on the trivial $\mathbb{R}^m$-bundle onto $TM$: for

a vector field $U$ and tangent vector $v \in T_xM$ the covariant derivative of $U$ in the direction $v$ is given by

$$\check{\nabla}_v U = X(x)(d[y \mapsto Y_y U(y)]_x(v)). \tag{7}$$

It follows from Narasimhan & Ramanan's theory of universal connections that every metric connection on $TM$ can be obtained by a suitable choice of $X$, see [22], or [56] for a direct proof. To obtain the Levi-Civita connection we can use Nash's theorem to take an isometric embedding $j\colon M \to \mathbb{R}^m$ for some $m$ and then set $X(x) = (dj)^*_x$, the adjoint of $(dj)_x$. With $A = 0$ the resulting 'gradient' SDE has Brownian motions as solutions as required. For Riemannian symmetric spaces it may be useful to use the homogeneous space structure; for example if $M$ is a compact Lie group with bi-invariant metric we may take $\mathbb{R}^m$ to be a copy of the direct sum $\mathfrak{g} \oplus \mathfrak{g}$ of the Lie algebra, $\mathfrak{g} = T_{\mathrm{Id}}M$, of $M$ with itself and define $X(x)(e, e') = TR_x(e) - TL_x(e')$ with $A = 0$, where $TR_x$ and $TL_x$ are the derivatives of left and right translation by $x$, [22].

One basic result, extending estimates in [4], which contrasts with the stochastic development is the following.

**Theorem 2.1** ([27]). *Suppose the connection $\check{\nabla}$ induced by the SDE is the Levi-Civita connection. The the pull back by $\mathcal{I}$ of cylindrical one-forms on $C_{x_0}$ extends to a continuous linear map $\mathcal{I}^*\colon L^2\mathcal{H}^* \to L^2(C_0([0,T];\mathbb{R}^m); H^*)$ of $L^2$ $H$-one-forms on $C_{x_0}$ to those on the flat path space.*

Here for a cylindrical, or other one form, $\phi$, on $C_{x_0}$, the pull-back $H$-form $\mathcal{I}^*(\phi)$ is given by $\mathcal{I}^*(\phi)_\omega(h) = \phi(T_\omega\mathcal{I}(h))$ for $h \in H$. However, in general the $H$-derivative $T\mathcal{I}$ does not map $H$ into the Bismut tangent spaces and so for $H$-one-forms $\phi$ the pullback does not have a classical meaning, though it does have an expression as an Itô integral under our condition on $\check{\nabla}$. If $\check{\nabla}$ were not the Levi-Civita connection this integral would be a Skorohod integral with a consequent loss of differentiability expected, as for the stochastic development map in [49]. There is an important equivalent dual, or 'co-joint', version to this result. For this suppose $\alpha\colon C_{x_0} \to H$ is an $H$-vector field in $L^2$. For almost all $\sigma \in C_{x_0}$ we can 'integrate over the fibre of $\mathcal{I}$' at $\sigma$ to obtain $\overline{T\mathcal{I}(\alpha)}_\sigma \in T_\sigma C_{x_0}$. Mathematically that is achieved by taking the conditional expectation with respect to the $\sigma$-algebra $\mathcal{F}^{x_0}$ on $C_{x_0}$ generated by $\mathcal{I}$:

$$\overline{T\mathcal{I}(\alpha)}_\sigma = \mathbb{E}\{T\mathcal{I}(\alpha(-)) \mid \mathcal{I}(-) = \sigma\}.$$

**Theorem 2.2** ([27]). *Suppose the connection $\check{\nabla}$ induced by the SDE is the Levi-Civita connection. Then for all $H$-vector fields $\alpha$ in $L^2$, we have $\overline{T\mathcal{I}(\alpha)}_\sigma \in \mathcal{H}_\sigma$ almost surely, giving a continuous linear map $\overline{T\mathcal{I}(-)}\colon L^2(C_{x_0}; H) \to L^2\mathcal{H}$.*

When $\alpha$ is constant, with value $h$ say, we write $\overline{TI}_\sigma(h)$ for $\overline{T\mathcal{I}(\alpha)}_\sigma$. This map was known earlier, [22], to map $H$ isomorphically onto $\mathcal{H}_\sigma$, with our assumption on $\check{\nabla}$. In fact it has the explicit expression

$$\overline{TI}_\sigma(h)_t = \boldsymbol{W}(X(\sigma(-))\dot{h}), \tag{8}$$

where $\boldsymbol{W}\colon L^2([0,T];TM) \to \mathcal{H}$ is an isomorphism of the Bismut tangent 'bundle', where defined, with the $L^2$-tangent bundle $L^2TC_{x_0}$ of $C_{x_0}$ given by

$$L^2T_\sigma C_{x_0} = \Big\{ v\colon [0,T] \to TM \text{ such that } v(t) \in T_{\sigma(t)}M,\, 0 \leqslant t \leqslant T \text{ and } \int_0^T |v(t)|^2_{\sigma(t)}\, dt < \infty \Big\}.$$

The isomorphism is the inverse of the 'damped derivative' along the paths of $C_{x_0}$:

$$\frac{\mathbb{D}}{dt} = \frac{D}{dt} + \frac{1}{2}\,\mathrm{Ric}^{\#}\colon \mathcal{H} \to L^2TC_{x_0}, \tag{9}$$

where $\mathrm{Ric}^{\#}\colon TM \to TM$ corresponds to the Ricci curvature.

It is convenient to give $\mathcal{H}$ the Riemannian metric and bundle structure it inherits from this isomorphism with the bundle of $L^2$ 'tangent vectors'. The latter is a smooth Hilbert bundle over $C_{x_0}$ with structure group $C_{\mathrm{Id}}([0,T];O(n))$. It also has a natural metric, 'Levi-Civita', connection, the 'pointwise connection' induced from the Levi-Civita connection on $M$, [19]. Moving this to $\mathcal{H}$ by $\boldsymbol{W}$ gives a metric connection which is easily seen to be that projected onto $\mathcal{H}$ by $\overline{T\mathcal{I}}$, in the same way as we defined $\check{\nabla}$. This connection agrees with the 'damped Markovian' connection of Cruzeiro & Fang, see [9], referred to above. It can be used to define higher order derivative operators and Sobolev spaces, and Sobolev spaces of sections of $\mathcal{H}$, e.g. $\mathbb{D}^{2,1}\mathcal{H}$, the domain of the $L^2$-closure of the covariant $H$-derivative acting on sections of $\mathcal{H}$. The latter is shown to be in the domain of div in [27]: a result proved by M. & P. Kree for classical Wiener measure in 1983.

We can define an $L^2$-function $f\colon C_{x_0} \to \mathbb{R}$ to be *weakly differentiable* if it is in the domain of the adjoint of the restriction of div to $\mathbb{D}^{2,1}\mathcal{H}$. Let $W^{2,1}$ denote the space of such functions with its graph norm. Thus for $f \in W^{2,1}$ there exists $\widetilde{df} \in L^2\mathcal{H}^*$ such that if $V \in \mathbb{D}^{2,1}\mathcal{H}$ then

$$\int_{C_{x_0}} f(\sigma)\,\mathrm{div}(V(\sigma))\,d\mu_{x_0} = -\int_{C_{x_0}} \widetilde{df}(V(\sigma))\,d\mu_{x_0}.$$

For paths on $\mathbb{R}^m$ it follows from [60] that weak differentiability implies differentiability, in our Sobolev sense.

We have the following intertwining result:

**Theorem 2.3** ([27]). *Suppose the connection $\check{\nabla}$ induced by the SDE is the Levi-Civita connection. Then $f \in W^{2,1}$ if and only if $f \circ \mathcal{I} \in \mathbb{D}^{2,1}C_0([0,T];\mathbb{R}^m)$ and composition with $\mathcal{I}$ gives a continuous linear map of $W^{2,1}$ onto the space $\mathbb{D}^{2,1}_{\mathcal{F}^{x_0}}$ of those elements in $\mathbb{D}^{2,1}C_0([0,T];\mathbb{R}^m)$ which are $\mathcal{F}^{x_0}$-measurable. Moreover for $f \in W^{2,1}$ we have*

$$d(f \circ \mathcal{I}) = \mathcal{I}^*\widetilde{df}.$$

Preliminary versions of some of the above results were given in [24]. A fundamental question is whether $W^{2,1} = \mathbb{D}^{2,1}$. Applying results of Eberle, [15], it is shown in [27] that this equality holds if and only if *Markov uniqueness* holds for the operator $\mathcal{L}$ defined above but with domain $\mathrm{Cyl}(M)$. Markov uniqueness is a weaker notion than essential self-adjointness. Probabilistically it relates to uniqueness of solutions to the martingale problem, and it essentially means that there is a unique extension which generates a Markov semigroup. Equality would also imply the independence of $\mathbb{D}^{2,1}$ from the choice of initial domain $\mathrm{Dom}(d_H)$. We do not know of any non-flat manifolds $M$ for which an answer is known to these questions. A positive answer would follow from a positive answer to the following.

*If* $f \in \mathbb{D}^{2,1}C_0([0,T];\mathbb{R}^m)$*, is its conditional expectation* $\mathbb{E}\{f|\mathcal{F}^{x_0}\}$ *also in* $\mathbb{D}^{2,1}$?

This is described concisely in [26], and in detail in [27], describing some partial results and correcting claims made in our 2004 Comptes-Rendues note. A discussion somewhat related to the above question, by Airault, Malliavin & Ren, is in [5].

**2.3. More general diffusion measures.** Let $\mathcal{A}$ be a smooth diffusion generator on $M$ i.e. it is a semi-elliptic second order differential operator with no zero order term, acting on real valued functions on $M$. Essentially as for the case $\mathcal{A} = \frac{1}{2}\Delta$, there is an induced measure $\mu_{x_0}^{\mathcal{A}}$ on $C_{x_0}$.

To extend the previous results to do analysis with such a measure we will suppose the principal symbol $\sigma^{\mathcal{A}}\colon T^*M \to TM$ of $\mathcal{A}$ has constant rank, and so has image in a sub-bundle $E$ of $TM$. This is equivalent to requiring that $\mathcal{A}$ has a Hormander form, as equation (6), with the vector fields $X^j$ being sections of $E$.

In general there is now no obvious choice of a connection with which to define 'Bismut tangent' spaces. We therefore choose any metric connection on $E$ and as before, using Narasimhan & Ramanan's theorem, take a stochastic differential equation (5) for which the induced connection $\breve{\nabla}$ on $E$ is that chosen one, and for which (6) holds. To define the Bismut tangent spaces it is convenient to use the *adjoint semi-connection*, $\widehat{\nabla}$, which allows differentiation of all smooth vector fields, but only in $E$-directions. It is defined by

$$\widehat{\nabla}_{U(x)}V = \breve{\nabla}_{V(x)}U + [U,V](x) \in T_xM$$

for $U$ a section of $E$ and $V$ a vector field on $M$, [22].

Adjoint connections were used in a similar way in order to use different Bismut tangent spaces for Brownian measures, by Driver in [12]. The adjoint of the Levi-Civita connection is itself; that of the flat left invariant connection on a Lie group is a flat right invariant connection. For more examples see [22]. Semi-connections are also called 'partial connections' or '$E$-connections'.

We now define $\mathcal{H}_\sigma$ to be the set of those $v \in T_\sigma C_{x_0}$ for which $\frac{\hat{\mathbb{D}}}{dt}(v) \in L^2([0,T];E)$ where

$$\frac{\hat{\mathbb{D}}}{dt} = \frac{\hat{D}}{dt} + \frac{1}{2}\,\check{\mathrm{Ric}}^{\#} - \breve{\nabla}_{-}A \tag{10}$$

where the covariant differentiation is done using the semi-connection while $\check{\mathrm{Ric}}^{\#}\colon TM \to E$ corresponds to the Ricci curvature for $\check{\nabla}$. If $A$ does not take values in $E$ then this operator needs special interpretation, [22]. Since $L^2([0,T];E)\bigcap L^2TC_{x_0}$ is a smooth Hilbert bundle, as for the case $E = TM$, with pointwise connection induced from $\check{\nabla}$, we can induce all this structure, at least almost surely, on $\mathcal{H}$. When there is a metric on $TM$ to which the semi-connection $\widehat{\nabla}$ is adapted, the theory goes essentially as before, [27]. If not there may be some loss of integrability in the intertwining, for example, but the operator $\mathcal{L}$ has a spectral gap; indeed there is a Log Sobolev inequality, [22]. The Dirichlet forms which arise in this situation are discussed in [28].

## 3. Towards an $L^2$-deRham–Hodge–Kodaira theory

**3.1. The spaces of $H$-forms.** Following Shigekawa's rather complete $L^2$-deRham theory for $H$-forms on abstract Wiener spaces it would be natural to base such a theory on sections of the dual bundles to the exterior products $\wedge^k\mathcal{H}$ of the Bismut tangent bundle, using the Hilbert space completion of the exterior powers of each $\mathcal{H}_\sigma$. However this runs into difficulties even at defining the exterior derivative of an $H$ one-form, $\phi$, say: Recall that the standard formula for the exterior derivative $d\phi$ is

$$d\phi\big(U(x)\wedge V(x)\big) = \mathcal{L}_U\big(\phi(V(-))\big)(x) - \mathcal{L}_V\big(\phi(U(-))\big)(x) - \phi\big([U,V](x)\big)$$

for vector fields $U$ and $V$. However if $U$ and $V$ are $H$-vector fields their bracket need not be and so if $\phi$ is an $H$-form the last term in the expression above will not in general be defined. One way round this is to interpret this final term as a stochastic integral, in general a Skorohod integral. This was carried through by Léandre in [47] where he obtained a deRham complex in this situation and for loop spaces, proving that the resulting deRham cohomology agrees with the topological real cohomology. However this was not really an $L^2$ theory and did not include a version of the Hodge–Kodaira Laplacian.

A proposal made in [24] was to modify the definitions of $k$-forms by replacing the spaces $\wedge^k\mathcal{H}$ by Hilbert spaces $\mathcal{H}^{(k)}$, for $k = 1, 2, \ldots$, continuously included in the projective exterior powers $\wedge^k TC_{x_0}$. For the 'projective exterior powers' the completion is made using the largest cross norm and the usual, geometric, differential forms are sections of the dual bundles $(\wedge^k TC_{x_0})^*$. Our $H$-$k$-forms will be sections of the dual bundles $\mathcal{H}^{(k)*}$.

To define $\mathcal{H}^{(k)}$, for simplicity we will deal only with the case of Brownian motion measures and Levi-Civita connections. The more general situation is touched upon in [25]; for details of the following see [23]. Take an SDE as in Section 2.2 with corresponding Itô map $\mathcal{I}$. It is shown that the map $h_1\wedge\cdots\wedge h_k \mapsto T_\omega\mathcal{I}(h_1)\wedge\cdots\wedge T_\omega(h_k)$ determines a continuous linear map $\wedge^k T_\omega\mathcal{I}\colon \wedge^k H \to \wedge^k T_{\mathcal{I}(\omega)}C_{x_0}$ from Hilbert space to Banach space. As done in Section 2.2 integrate over the fibres

of $\mathcal{I}$ to define

$$\overline{\wedge T\mathcal{I}}_\sigma : \wedge^k H \to \wedge^k T_\sigma C_{x_0}$$

for almost all $\sigma \in C_{x_0}$, by the conditional expectation:

$$\overline{\wedge T\mathcal{I}}_\sigma(\underline{h}) = \mathbb{E}\{\wedge^k T\mathcal{I}(\underline{h})|\mathcal{I} = \sigma\}$$

for $\underline{h} \in \wedge^k H$. We then let $\mathcal{H}_\sigma^{(k)}$ be the image of $\overline{\wedge T\mathcal{I}}_\sigma$ with its quotient Hilbert space structure. Thus $\mathcal{H}^{(1)} = \mathcal{H}$. As with the case $k = 1$ these spaces depend only on the Riemannian structure of $M$, not on the choice of SDE we used to construct them (provided $\breve{\nabla}$ is the Levi-Civita connection).

For $k = 2$ there is a detailed description. For this let $\boldsymbol{R}: \wedge^2 TC_{x_0} \to \mathbb{L}(\mathcal{H}; \mathcal{H})$ be the curvature operator of the damped Markovian connection on $\mathcal{H}$, see Section 2.2, and let $\mathbb{T}: \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ be its torsion. We have

$$\mathcal{H}^{(2)} = \{U \in \wedge^2 TC_{x_0} : U - \boldsymbol{R}(U) \in \wedge^2 \mathcal{H}\}$$

with inner product having the norm $|U|_{\mathcal{H}^2} = |U - \boldsymbol{R}(U)|_{\wedge^2\mathcal{H}}$. Alternatively, inverting $\mathrm{Id} - \boldsymbol{R}$, we have

$$\mathcal{H}^{(2)} = \{V + \boldsymbol{Q}(V) : V \in \wedge^2 \mathcal{H}\}$$

where the linear map $\boldsymbol{Q}$ can be expressed in terms of the curvature of $M$ and involves a 'damped translation' of 2-vectors on $M$ where the damping is by the second Weitzenbock curvature, just as the first, the Ricci curvature, appears in equation (9). It turns out that '$\mathrm{div}$' $\boldsymbol{Q}(u \wedge v) = \frac{1}{2}\mathbb{T}(u, v)$ for any bounded *adapted* $H$-vector fields $u$ and $v$ in the sense that for any smooth cylindrical one-form $\phi$ on $C_{x_0}$ we have

$$\int_{C_{x_0}} d\phi(\boldsymbol{Q}(u \wedge v)d\mu_{x_0} = -\frac{1}{2}\int_{C_{x_0}} \phi(\mathbb{T}(u, v))d\mu_{x_0}.$$

This relates to a result of Cruzeiro–Fang, [10], that for suitable $u$ and $v$ the torsion $\mathbb{T}(u, v)$ has 'divergence' zero, in the corresponding sense.

If we define the exterior $H$-derivative as usual on cylindrical one forms $\phi$ but restrict the resulting $(d_H\phi)_\sigma : \wedge^2 T_\sigma C_{x_0} \to \mathbb{R}$ to $\mathcal{H}_\sigma^{(2)}$ we obtain a map, with domain the smooth cylindrical one forms, into the $L^2$ $H$ two-forms, $L^2\ \mathcal{H}^{(2)*}$. The cylindrical one forms when restricted to $\mathcal{H}^*$ form a dense subspace of $L^2\mathcal{H}^*$ and it turns out that this map is closable as an operator on $L^2\mathcal{H}^*$. We obtain a closed exterior derivative operator

$$d^1 : \mathrm{Dom}(d) \subset L^2\mathcal{H}^* \to L^2\mathcal{H}^{(2)*}$$

with a dual operator $\mathrm{div}: \mathrm{Dom}(\mathrm{div}) \subset L^2\mathcal{H}^{(2)} \to L^2\mathcal{H}$.

The covariant derivative determined by the damped Markovian connection on $\mathcal{H}$ can be considered as a closed operator $\nabla$ from its domain, $\mathbb{D}^{2,1}\mathcal{H}$, in $L^2\mathcal{H}$ to $L^2(\mathcal{H} \otimes \mathcal{H})$ and so has an adjoint $\nabla^*$. The following suggests that our construction is a natural one, but the condition of adaptedness on the vector fields is essential:

**Proposition 3.1** ([23]). *Let $u$ and $v$ be bounded and adapted $H$-vector fields on $C_{x_0}$. Suppose $u, v \in \mathbb{D}^{2,1}\mathcal{H}$ then $u \wedge v \in \mathrm{Dom}\,\nabla^*$ and*

$$\nabla^*(u \wedge v) = \mathrm{div}((\mathrm{Id} + \boldsymbol{Q})(u \wedge v)).$$

It turns out that the exterior product $\phi^1 \wedge \phi^2$ of two $H$-one-forms can be considered as an $H$-two-form in a consistent way. Essentially this is because although an element in some $\mathcal{H}_\sigma^{(2)}$ is not in $\mathcal{H}_\sigma \otimes \mathcal{H}_\sigma$, a space which can be identified with the Hilbert–Schmidt maps on $\mathcal{H}_\sigma$, it can be identified with a bounded linear map on $\mathcal{H}_\sigma$, and elements of the uncompleted tensor product of $\mathcal{H}_\sigma$ with itself act as linear functionals on the bounded linear maps. We have then:

**Proposition 3.2** ([23]). *Suppose $f \in \mathbb{D}^{2,1}(C_{x_0}; \mathbb{R})$ and $\phi$ is a bounded $H$-one-form which is in the domain of the exterior derivative, and is bounded together with $d\phi$. Then $f\phi$ is in the domain of the exterior derivative and*

$$d^1(f\phi) = df \wedge \phi + f d^1\phi.$$

**3.2. A Hodge–Kodaira decomposition for one and two forms.** The key step to prove closability of the exterior derivative on these $H$-$k$-forms is to prove an analogue of Theorem 2.2. We would like a rich set of $L^2$ maps $\underline{h}\colon C_{x_0} \to \wedge^k\mathcal{H}$ such that

$$\overline{\wedge^k T\mathcal{I}(\underline{h})}_\sigma := \mathbb{E}\{\wedge^k T\mathcal{I}(\underline{h}) | \mathcal{I} = \sigma\} \in \mathcal{H}_\sigma^{(k)}$$

almost surely. For $k = 1$ this holds for all such $\underline{h}$ by Theorem 2.2. For $k = 2$ it is claimed for an adequately rich family in [23], for all relevant Itô maps, and for all $\underline{h}$ if the Itô map is defined via a symmetric space structure. It is unknown for higher $k$ largely because of the apparently complicated algebraic structure of the spaces $\mathcal{H}^{(k)}$ for higher $k$. (On the other hand in [25] it is shown that an important class of $k$-vector fields, defined for $k = 1, \ldots, n-1$ are $L^2$ sections of $\mathcal{H}^{(k)}$ when $k = 1, 2$: these are important in the sense that they give integration by parts results, or generalised 'Bismut-formulae', for the finite dimensional exterior derivatives $dP_t\phi$ of the heat semigroup on forms on $M$ in terms of a path integral of $\phi$ itself.)

From these results for $k = 1, 2$ we have now closed operators

$$d^k\colon \mathrm{Dom}(d^k) \subset L^2\Gamma\mathcal{H}^{(k)} \to L^2\Gamma\mathcal{H}^{(k+1)}$$

for $k = 1, 2$ with $d^2d^1 = 0$. This leads to the Hodge–Kodaira decomposition:

$$L^2\Gamma\mathcal{H}^k = \overline{\mathrm{Im}(d^{(k-1)})} \oplus \overline{\mathrm{Im}((d^k)^*)} \oplus (\ker d^k \cap \ker(d^{(k-1)})^*) \tag{11}$$

for $k = 1, 2$, as given for $k = 1$ in [24], and for $k = 2$ in [23]. Here $d^0$ refers to $d$, and in the case $k = 1$ the image of $d$ is closed by Fang's theorem, e.g. see the Clark–Ocone formula in [22]. Moreover we have self-adjoint operators $(d^k)^*d^k + d^{(k-1)}(d^{(k-1)})^*$ acting on the spaces of $L^2$ $H$-$k$-forms for $k = 1, 2$. For $k = 1$ the decomposition plus Fang's theorem shows that the space of $L^2$ harmonic one-forms represents the $L^2$-deRham cohomology group of $H$-one-forms.

**3.3. Lie groups with flat connection.** At present we have no information about even the first $L^2$ deRham group for non-flat manifolds. However in [31], Fang & Franchi considered the case where $M$ is a compact Lie group $G$ with bi-invariant metric. For the Bismut tangent spaces coming from a right invariant flat connection, the natural Itô map to use is that of a left invariant SDE, $dx_t = TL_{x_t} \circ dB_t$, for $B_\cdot$ a Brownian motion on the Lie algebra $\mathfrak{g}$. There is no 'redundant noise' and the derivative of the Itô map maps the Cameron–Martin space into the Bismut tangent spaces, and its exterior powers onto those of the Bismut spaces. There is no problem with the definition of the exterior derivative and they showed that the Itô map can be used to transfer Shigekawa's results for classical Wiener space, determining a Hodge–Kodaira decomposition, and giving the vanishing of $L^2$ harmonic forms and so of the corresponding deRham cohomology groups.

## 4. Loop spaces

We have not extended the Itô map techniques described above in any systematic way to the case of loop spaces, (but see [1]), and here will only briefly describe the basic set up and some relevant results. The surveys [48] and [2] give more information and references. Special motivation for the development of analysis on these spaces has come from the loop space approach to index theorems, as in [7], and the Hohn–Stolz conjecture, [59]. However note that this theory is based on tangent spaces of vectors which are in some sense in $L^{2,1}$ and it is not clear that this is always what is relevant to some physical or topological situations, e.g. see [32].

On the space of based loops, or more generally on the spaces $C_{x_0,y_0}$ of continuous paths $\sigma : [0, T] \to M$ with $\sigma(0) = x_0$ and $\sigma(1) = y_0$, for $y_0 \in M$, a natural measure to take is the *Brownian Bridge measure*, $\mu_{x_0,y_0}$, obtained by conditioning Brownian motion from $x_0$ to be at $y_0$ at time $T$. If equation (5) has solutions which are Brownian motions then the equation:

$$db_t = X(b_t) \circ dB_t + A(b_t)dt + \nabla \log p_{T-t}(b_t; y_0)dt \tag{12}$$

will have Itô map which sends Wiener measure to $\mu_{x_0,y_0}$. Here $p_t(x, y)$ is the heat kernel as in Section 2.1.

For the space $L(M)$ of free loops, i.e. of continuous $\sigma : S^1 \to M$, there is Bismut's measure, $\mu_L$, which can be defined as $\int_M p_T(y, y)\mu_{y,y}\, dy$ with $T = 2\pi$, [7]. This measure is invariant under the action of $S^1$. A variant of this when $M$ is a Lie group is to average using normalised Haar measure rather than the heat kernel. Either of these loop spaces could be furnished with a *heat kernel measure*, $\mu_h$. This is defined by choosing a base point, e.g. a constant loop, and constructing a 'Brownian motion' on the loop space starting at that point, running it for some fixed time, $\tau$ say, and using its probability distribution as $\mu_h$. This will depend on $\tau$ and for free loops an extra averaging over the initial base point to retain $S^1$-invariance is needed. The

construction of such Brownian motions goes back to Baxendale, see [6] and Gaveau & Mazet, [34] but has been most developed for loop groups, [51]. For based paths on a compact simply connected Lie group this measure has been shown to be equivalent to the Brownian Bridge measure, [3].

In these contexts, by results of several people including M. P. & P. Malliavin, Driver, Hsu, Leandre, Enchev & Stroock, and Aida there are integration by parts formulae and associated Sobolev spaces based on Bismut tangent spaces defined similarly to those above, though for Lie groups flat connections are often used to define the Bismut tangent spaces. For the Brownian bridge and Bismut measures there are cohomology results: using stochastic Chen forms in [40], 'Sobolev differential forms' [47], and more recently 'Chen–Souriau cohomology' defined via a 'stochastic diffeology', see [48]. In general the resulting cohomology agrees with the usual singular real cohomology. We refer the reader to MathScinet to see the variety of constructions in these and related situations by R. Leandre.

For compact Lie groups with bi-invariant metrics, and with Bismut tangent spaces defined by flat left, or right, invariant connections, Fang & Franchi were able to extend their results for path spaces to based loops defining Hodge–Kodaira operators on forms and giving a 'Weitzenbock formulae' for them, [30]. This formula rather clearly shows the form of these operators as 'Witten' or 'Bismut' Laplacians where the 'perturbing' vector field is not an $H$-vector field, and gives rise to stochastic integrals in the formulae. The curvature part of the formulae has a Ricci term, which requires a careful summation, as in [32].

One striking result for the Brownian bridge measure is the following by Eberle:

**Theorem 4.1** ([16]). *Suppose the compact manifold $M$ has a closed geodesic for which there is a neighbourhood in $M$ of constant negative curvature. Then on the loop spaces $C_{x_0,x_0}$ with Brownian bridge measure, and $L(M)$ with Bismut measure, the self-adjoint operator $\mathcal{L} = -d^*d$ does not have a spectral gap.*

Spectral gaps for the Hodge–Kodaira 'Laplacians' are important in Hodge theory since they correspond to the (exterior) derivative operators having closed range in $L^2$. At present it is unknown if there is ever a spectral gap for $\mathcal{L}$ for these measures for loops on non-flat manifolds, e.g. on spheres. However for heat kernel measures on compact Lie groups with bi-invariant metrics Driver & Lohrenz proved the existence of a Log Sobolev inequality and so of a spectral gap for $\mathcal{L}$, see [29].

An alternative approach to based loops has been to represent them by 'submanifolds' of classical Wiener space by choosing a suitable (i.e. a quasi-continuous) version of the stochastic development and considering the inverse image $\widetilde{C}$, say, under it, of the based loops on $M$. This construction is dependent on 'quasi-sure' analysis, see [50], where our measure theoretic concepts are refined potential theoretically, so that $\widetilde{C}$ can be defined up to sets of capacity zero. To a certain extent this allows $\widetilde{C}$ to be treated as a submanifold of co-dimension the dimension of $M$, with a differential form theory and Weitzenbock formula, see [41], and [45].

## References


[1] Aida, Shigeki, Differential calculus on path and loop spaces. II. Irreducibility of Dirichlet forms on loop spaces. *Bull. Sci. Math.* **122** (8) (1998), 635–666.

[2] Aida, Shigeki, Stochastic analysis on loop spaces. *Sugaku Expositions* **13** (2) (2000), 197–214.

[3] Aida, Shigeki, and Driver, Bruce K., Equivalence of heat kernel measure and pinned Wiener measure on loop groups. *C. R. Acad. Sci. Paris Sér. I Math.* **331** (9) (2000), 709–712.

[4] Aida, Shigeki, and Elworthy, David, Differential calculus on path and loop spaces. I. Logarithmic Sobolev inequalities on path spaces. *C. R. Acad. Sci. Paris Sér. I Math.* **321** (1) (1995), 97–102.

[5] Airault, Hélène , Malliavin, Paul, and Ren, Jiagang, Geometry of foliations on the Wiener space and stochastic calculus of variations. *C. R. Math. Acad. Sci. Paris* **339** (9) (2004), 637–642.

[6] Baxendale, Peter, Wiener processes on manifolds of maps. *Proc. Roy. Soc. Edinburgh Sect. A* **87** (1–2) (1980/81), 127–152.

[7] Bismut, Jean-Michel, Index theorem and equivariant cohomology on the loop space. *Comm. Math. Phys.* **98** (2) (1985), 213–237.

[8] Cameron, R. H., and Martin, W. T., The transformation of Wiener integrals by nonlinear transformations. *Trans. Amer. Math. Soc.* **66** (1949), 253–283.

[9] Cruzeiro, Ana Bela, and Fang, Shizan, An $L^2$ estimate for Riemannian anticipative stochastic integrals. *J. Funct. Anal.* **143** (2) (1997), 400–414.

[10] Cruzeiro, Ana Bela, and Fang, Shizan, Weak Levi-Civita connection for the damped metric on the Riemannian path space and vanishing of Ricci tensor in adapted differential geometry. *J. Funct. Anal.* **185** (2) (2001), 681–698.

[11] Cruzeiro, Ana Bela, and Malliavin, Paul, Renormalized differential geometry on path space: structural equation, curvature. *J. Funct. Anal.* **139** (1) (1996), 119–181.

[12] Driver, B. K., A Cameron-Martin type quasi-invariance theorem for Brownian motion on a compact Riemannian manifold. *J. Funct. Anal.* **100** (1992), 272–377.

[13] Driver, Bruce K., and Röckner, Michael, Construction of diffusions on path and loop spaces of compact Riemannian manifolds. *C. R. Acad. Sci. Paris Sér. I Math.* **315** (5) (1992), 603–608.

[14] Dudley, R. M., Feldman, Jacob, and Le Cam, L., On seminorms and probabilities, and abstract Wiener spaces. *Ann. of Math.* (2) **93** (1971), 390–408.

[15] Eberle, Andreas, *Uniqueness and non-uniqueness of semigroups generated by singular diffusion operators*. Lecture Notes in Math. 1718, Springer-Verlag, Berlin 1999.

[16] Eberle, Andreas, Absence of spectral gaps on a class of loop spaces. *J. Math. Pures Appl.* (9) **81** (10) (2002), 915–955.

[17] Eells, J., and Elworthy, K. D., On Fredholm manifolds. In *Actes du Congrès International des Mathématiciens* (Nice, 1970), Tome 2, Gauthier-Villars, Paris 1971, 215–219.

[18] Eells, J., and Elworthy, K. D., Wiener integration on certain manifolds. In *Problems in non-linear analysis* (C.I.M.E., IV Ciclo, Varenna, 1970), Edizioni Cremonese, Rome 1971, 67–94.

[19] Eliasson, H., Geometry of manifolds of maps. *J. Differential Geom.* **1** (1967), 169–194.

[20] Elworthy, K. D., Gaussian measures on Banach spaces and manifolds. In *Global analysis and its applications* (Lectures, Internat. Sem. Course, Internat. Centre Theoret. Phys., Trieste, 1972), Vol. II, Internat. Atomic Energy Agency, Vienna 1974, 151–166.

[21] Elworthy, K. D., Differential invariants of measures on Banach spaces. In *Vector space measures and applications* (Proc. Conf., Univ. Dublin, Dublin 1977), II, Lecture Notes in Math. 644, Springer-Verlag, Berlin 1978, 159–187.

[22] Elworthy, K. D., LeJan, Y., and Li, Xue-Mei, *On the geometry of diffusion operators and stochastic flows*. Lecture Notes in Math. 1720, Springer-Verlag, Berlin 1999.

[23] Elworthy, K. D., and Li, Xue-Mei, An $L^2$ theory for differential forms on path spaces. In preparation.

[24] Elworthy, K. D., and Li, Xue-Mei, Special Itô maps and an $L^2$ Hodge theory for one forms on path spaces. In *Stochastic processes, physics and geometry: new interplays, I* (Leipzig, 1999), CMS Conf. Proc. 28, Amer. Math. Soc., Providence, RI, 2000, 145–162.

[25] Elworthy, K. D., and Li, Xue-Mei, Some families of $q$-vector fields on path spaces. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* **6** (suppl.) (2003), 1–27.

[26] Elworthy, K. D., and Li, Xue-Mei, Intertwining and the Markov uniqueness problem on path spaces. In *Stochastic Partial Differential Equations and Applications VII*, Lecture Notes in Pure and Applied Mathematics 245, Chapman and Hall/CRC, Boca Raton, FL, 2006, 89–95.

[27] Elworthy, K. D., and Li, Xue-Mei, Ito maps and analysis on path spaces. Warwick Preprint, also www.xuemei.org, 2005.

[28] Elworthy, K. David, and Ma, Zhi-Ming, Vector fields on mapping spaces related Dirichlet forms and diffusions. *Osaka J. Math.* **34** (3) (1997), 629–651.

[29] Fang, Shizan, Integration by parts formula and logarithmic Sobolev inequality on the path space over loop groups. *Ann. Probab.* **27** (2) (1999), 664–683.

[30] Fang, Shizan, and Franchi, Jacques, De Rham-Hodge-Kodaira operator on loop groups. *J. Funct. Anal.* **148** (2) (1997), 391–407.

[31] Fang, Shizan, and Franchi, Jacques, A differentiable isomorphism between Wiener space and path group. In *Séminaire de Probabilités, XXXI*, Lecture Notes in Math. 1655, Springer-Verlag, Berlin 1997, 54–61.

[32] Freed, Daniel S., The geometry of loop groups. *J. Differential Geom.* **28** (2) (1988), 223–276.

[33] Gangolli, Ramesh, On the construction of certain diffusions on a differentiable manifold. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 2 (1964), 406–419.

[34] Gaveau, Bernard, and Mazet, Edmond, Diffusion et intégration sur les espaces de lacets. *C. R. Acad. Sci. Paris Sér. A-B* **289** (13) (1979), A643–A646.

[35] Gross, Leonard, Potential theory on Hilbert space. *J. Funct. Anal.* **1** (1967), 123–181.

[36] Gross, Leonard, Abstract Wiener measure and infinite dimensional potential theory. In *Lectures in Modern Analysis and Applications, II*, Lecture Notes in Math. 140. Springer-Verlag, Berlin 1970, 84–116.

[37] Hsu, Elton P., *Stochastic analysis on manifolds*. Graduate Studies in Mathematics 38, Amer. Math. Soc., Providence, RI, 2002.

[38] Ikeda, N., and Watanabe, S., *Stochastic Differential Equations and Diffusion Processes*. Second edition, North-Holland, Kodansha Ltd., Amsterdam, Tokyo 1989.

[39] Itô, Kiyosi, A measure-theoretic approach to Malliavin calculus. In *New trends in stochastic analysis (Charingworth, 1994)*, pages 220–287. World Sci. Publishing, River Edge, NJ, 1997.

[40] Jones, J. D. S., and Léandre, R., $L^p$-Chen forms on loop spaces. In *Stochastic analysis* (Durham, 1990), London Math. Soc. Lecture Note Ser. 167, Cambridge University Press, Cambridge, 1991, 103–162.

[41] Kazumi, Tetsuya, and Shigekawa, Ichiro, Differential calculus on a submanifold of an abstract Wiener space. II. Weitzenböck formula. In *Dirichlet forms and stochastic processes* (Beijing, 1993), de Gruyter, Berlin 1995, 235–251.

[42] Kuo, Hui Hsiung, Integration theory on infinite-dimensional manifolds. *Trans. Amer. Math. Soc.* **159** (1971), 57–78.

[43] Kuo, Hui Hsiung, Diffusion and Brownian motion on infinite-dimensional manifolds. *Trans. Amer. Math. Soc.* **169** (1972), 439–459.

[44] Kusuoka, Shigeo, Degree theorem in certain Wiener Riemannian manifolds. In *Stochastic analysis* (Paris, 1987), Lecture Notes in Math. 1322, Springer-Verlag, Berlin 1988, 93–108.

[45] Kusuoka, Shigeo, de Rham cohomology of Wiener-Riemannian manifolds. In *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. II, Math. Soc. Japan., Tokyo 1991, 1075–1082.

[46] Kusuoka, Shigeo, Analysis on Wiener spaces. II. Differential forms. *J. Funct. Anal.* **103** (2) (1992), 229–274.

[47] Léandre, R., Cohomologie de Bismut-Nualart-Pardoux et cohomologie de Hochschild entière. In *Séminaire de Probabilités, XXX*, Lecture Notes in Math. 1626, Springer-Verlag, Berlin 1996, 68–99.

[48] Léandre, R., Analysis on loop spaces, and topology. *Mat. Zametki* **72** (2) (2002), 236–257.

[49] Li, Xiang-Dong, Sobolev spaces and capacities theory on path spaces over a compact Riemannian manifold. *Probab. Theory Relat. Fields* **125** (2003), 96–134.

[50] Malliavin, P., *Stochastic analysis*. Grundlehren Math. Wiss. 313, Springer-Verlag, Berlin 1997.

[51] Malliavin, P., Diffusion on the loops. In *Conference on harmonic analysis in honor of Antoni Zygmund* (Chicago, Ill., 1981), Vol. II, Wadsworth Math. Ser., Wadsworth, Belmont, CA, 1983, 764–782.

[52] Norris, J. R., Twisted sheets. *J. Funct. Anal.* **132** (2) (1995), 273–334.

[53] Nualart, David, *The Malliavin calculus and related topics*. Probab. Appl. (N.Y.), Springer-Verlag, New York 1995.

[54] Palais, Richard S., *Foundations of global non-linear analysis*. Mathematics Lecture Note Series, W. A. Benjamin, Inc., New York, Amsterdam 1968.

[55] Ann Piech, M., A model for an infinite-dimensional Laplace-Beltrami operator. *Indiana Univ. Math. J.* **31** (3) (1982), 327–340.

[56] Quillen, D., Superconnections; character forms and the Cayley transform. *Topology* **27** (2) (1988), 211–238.

[57] Segal, Irving, Algebraic integration theory. *Bull. Amer. Math. Soc.* **71** (1965), 419–489.

[58] Shigekawa, Ichirō, de Rham-Hodge-Kodaira's decomposition on an abstract Wiener space. *J. Math. Kyoto Univ.* **26** (2) (1986), 191–202.

[59] Stolz, Stephan, A conjecture concerning positive Ricci curvature and the Witten genus. *Math. Ann.* **304** (4) (1996), 785–800.

[60] Sugita, H., On a characterization of the Sobolev spaces over an abstract wiener space. *J. Math. Kyoto Univ.* **25** (4) (1985), 717–725.

[61] Üstünel, A. Süleyman, and Zakai, Moshe, *Transformation of measure on Wiener space*. Springer Monogr. Math., Springer-Verlag, Berlin 2000.

[62] Yamasaki, Y., *Measures on infinite-dimensional spaces*. Series in Pure Mathematics 5, World Scientific Publishing Co., Singapore 1985.

Mathematics Institute, Warwick University, Coventry CV4 7AL, U.K.

E-mail: kde@maths.warwick.ac.uk

Department of Mathematical Sciences, Loughborough University, Loughborough, LE11 3TU, U.K.

E-mail: xue-mei.li@lboro.ac.uk

# Statistical challenges with high dimensionality: feature selection in knowledge discovery

Jianqing Fan and Runze Li*

**Abstract.** Technological innovations have revolutionized the process of scientific research and knowledge discovery. The availability of massive data and challenges from frontiers of research and development have reshaped statistical thinking, data analysis and theoretical studies. The challenges of high-dimensionality arise in diverse fields of sciences and the humanities, ranging from computational biology and health studies to financial engineering and risk management. In all of these fields, variable selection and feature extraction are crucial for knowledge discovery. We first give a comprehensive overview of statistical challenges with high dimensionality in these diverse disciplines. We then approach the problem of variable selection and feature extraction using a unified framework: penalized likelihood methods. Issues relevant to the choice of penalty functions are addressed. We demonstrate that for a host of statistical problems, as long as the dimensionality is not excessively large, we can estimate the model parameters as well as if the best model is known in advance. The persistence property in risk minimization is also addressed. The applicability of such a theory and method to diverse statistical problems is demonstrated. Other related problems with high-dimensionality are also discussed.

**Mathematics Subject Classification (2000).** Primary 62J99; Secondary 62F12.

**Keywords.** AIC, BIC, LASSO, bioinformatics, financial econometrics, model selection, oracle property, penalized likelihood, persistent, SCAD, statistical learning.

## 1. Introduction

Technological innovations have had deep impact on society and on scientific research. They allow us to collect massive amount of data with relatively low cost. Observations with curves, images or movies, along with many other variables, are frequently seen in contemporary scientific research and technological development. For example, in biomedical studies, huge numbers of magnetic resonance images (MRI) and functional MRI data are collected for each subject with hundreds of subjects involved. Satellite imagery has been used in natural resource discovery and agriculture, collecting thousands of high resolution images. Examples of these kinds are plentiful in computational biology, climatology, geology, neurology, health science, economics,

*Fan's research was supported partially by NSF grant DMS-0354223, DMS-0532370 and NIH R01-GM072611. Li's research was supported by NSF grant DMS-0348869 and National Institute on Drug Abuse grant P50 DA10075. The authors would like to thank Professors Peter Hall and Michael Korosok for their constructive comments and John Dziak for his assistance. The article was presented by Jianqing Fan.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

and finance among others. Frontiers of science, engineering and the humanities differ in the problems of their concerns, but nevertheless share one common theme: massive and high-throughput data have been collected and new knowledge needs to be discovered using these data. These massive collections of data along with many new scientific problems create golden opportunities and significant challenges for the development of mathematical sciences.

The availability of massive data along with new scientific problems have reshaped statistical thinking and data analysis. Dimensionality reduction and feature extraction play pivotal roles in all high-dimensional mathematical problems. The intensive computation inherent in these problems has altered the course of methodological development. At the same time, high-dimensionality has significantly challenged traditional statistical theory. Many new insights need to be unveiled and many new phenomena need to be discovered. There is little doubt that the high dimensional data analysis will be the most important research topic in statistics in the 21st century [19].

Variable selection and feature extraction are fundamental to knowledge discovery from massive data. Many variable selection criteria have been proposed in the literature. Parsimonious models are always desirable as they provide simple and interpretable relations among scientific variables in addition to reducing forecasting errors. Traditional variable selection such as $C_p$, AIC and BIC involves a combinatorial optimization problem, which is NP-hard, with computational time increasing exponentially with the dimensionality. The expensive computational cost makes traditional procedures infeasible for high-dimensional data analysis. Clearly, innovative variable selection procedures are needed to cope with high-dimensionality.

Computational challenges from high-dimensional statistical endeavors forge cross-fertilizations among applied and computational mathematics, machine learning, and statistics. For example, Donoho and Elad [20] and Donoho and Huo [21] show that the NP-hard best subset regression can be solved by a penalized $L_1$ least-squares problem, which can be handled by a linear programming, when the solution is sufficiently sparse. Wavelets are widely used in statistics function estimation and signal processing [1], [14], [17], [23], [24], [64], [65], [71]. Algebraic statistics, the term coined by Pistone, Riccomagno, Wynn [73], uses polynomial algebra and combinatorial algorithms to solve computational problems in experimental design and discrete probability [73], conditional inferences based on Markovian chains [16], parametric inference for biological sequence analysis [72], and phylogenetic tree reconstruction [78].

In high-dimensional data mining, it is helpful to distinguish two types of statistical endeavors. In many machine learning problems such as tumor classifications based on microarray or proteomics data and asset allocations in finance, the interests often center around the classification errors, or returns and risks of selected portfolios rather than the accuracy of estimated parameters. On the other hand, in many other statistical problems, concise relationship among dependent and independent variables are needed. For example, in health studies, we need not only to identify risk factors, but also to assess accurately their risk contributions. These are needed for prognosis and understanding the relative importance of risk factors. Consistency results are

inadequate for assessing the uncertainty in parameter estimation. The distributions of selected and estimated parameters are needed. Yet, despite extensive studies in classical model selection techniques, no satisfactory solutions have yet been produced.

In this article, we address the issues of variable selection and feature extraction using a unified framework: penalized likelihood methods. This framework is applicable to both machine learning and statistical inference problems. In addition, it is applied to both exact and approximate statistical modeling. We outline, in Section 2, some high-dimensional problems from computational biology, biomedical studies, financial engineering, and machine learning, and then provide a unified framework to address the issues of feature selection in Sections 3 and 4. In Sections 5 and 6, the framework is then applied to provide solutions to some problems outlined in Section 2.

## 2. Challenges from sciences and humanities

We now outline a few problems from various frontiers of research to illustrate the challenges of high-dimensionality. Some solutions to these problems will be provided in Section 6.

**2.1. Computational biology.** Bioinformatic tools have been widely applied to genomics, proteomics, gene networks, structure prediction, disease diagnosis and drug design. The breakthroughs in biomedical imaging technology allow scientists to monitor large amounts of diverse information on genetic variation, gene and protein functions, interactions in regulatory processes and biochemical pathways. Such technology has also been widely used for studying neuron activities and networks. Genomic sequence analysis permits us to understand the homologies among different species and infer their biological structures and functionalities. Analysis of the network structure of protein can predict the protein biological function. These quantitative biological problems raise many new statistical and computational problems. Let us focus specifically on the analysis of microarray data to illustrate some challenges with dimensionality.

DNA microarrays have been widely used in simultaneously monitoring mRNA expressions of thousands of genes in many areas of biomedical research. There are two popularly-used techniques: c-DNA microarrays [5] and Affymetrix GeneChip arrays [61]. The former measures the abundance of mRNA expressions by mixing mRNAs of treatment and control cells or tissues, hybridizing with cDNA on the chip. The latter uses combined intensity information from 11-20 probes interrogating a part of the DNA sequence of a gene, measuring separately mRNA expressions of treatment and control cells or tissues. Let us focus further on the cDNA microarray data.

The first statistical challenge is to remove systematic biases due to experiment variations such as intensity effect in the scanning process, block effect, dye effect, batch effect, amount of mRNA, DNA concentration on arrays, among others. This is collectively referred to as normalization in the literature. Normalization is critical

for multiple array comparisons. Statistical models are needed for estimation of these systematic biases in presence of high-dimensional nuisance parameters from treatment effects on genes. See, for example, lowess normalization in [26], [83], semiparametric model-based normalization by [36], [37], [50], and robust normalization in [63]. The number of significantly expressed genes is relatively small. Hence, model selection techniques can be used to exploit the sparsity. In Section 6.1, we briefly introduce semiparametric modeling techniques to issues of normalization of cDNA microarray.

Once systematic biases have been removed, the statistical challenge becomes selecting statistically significant genes based on a relatively small sample size of arrays (e.g. $n = 4, 6, 8$). Various testing procedures have been proposed in the literature. See, for example, [30], [37], [50], [83], [84]. In carrying out simultaneous testing of orders of hundreds or thousands of genes, classical methods of controlling the probability of making one falsely discovered gene are no longer relevant. Therefore various innovative methods have been proposed to control the false discovery rates. See, for example, [2], [22], [25], [27], [44], [57], [77]. The fundamental assumption in these developments is that the null distribution of test statistics can be determined accurately. This assumption is usually not granted in practice and new probabilistic challenge is to answer the questions how many simultaneous hypotheses can be tested before the accuracy of approximations of null distributions becomes poor. Large deviation theory [45], [46], [53] is expected to play a critical role in this endeavor. Some progress has been made using maximal inequalities [55].

Tumor classification and clustering based on microarray and proteomics data are another important class of challenging problems in computational biology. Here, hundreds or thousands of gene expressions are potential predictors, and the challenge is to select important genes for effective disease classification and clustering. See, for example, [79], [82], [88] for an overview and references therein.

Similar problems include time-course microarray experiments used to determine the expression pathways over time [79], [80] and genetic networks used for understanding interactions in regulatory processes and biochemical pathways [58]. Challenges of selecting significant genes over time and classifying patterns of gene expressions remain. In addition, understanding genetic network problems requires estimating a huge covariance matrix with some sparsity structure. We introduce a modified Cholesky decomposition technique for estimating large scale covariance matrices in Section 6.1.

**2.2. Health studies.** Many health studies are longitudinal: each subject is followed over a period of time and many covariates and responses of each subject are collected at different time points. Framingham Heart Study (FHS), initiated in 1948, is one of the most famous classic longitudinal studies. Documentation of its first 50 years can be found at the website of National Heart, Lung and Blood Institute (http://www.nhlbi.nih.gov/about/framingham/). One can learn more details about this study from the website of American Heart Association. In brief, the FHS follows a representative sample of 5,209 adult residents and their offspring aged 28–62 years in

Framingham, Massachusetts. These subjects have been tracked using (a) standardized biennial cardiovascular examination, (b) daily surveillance of hospital admissions, (c) death information and (d) information from physicians and other sources outside the clinic.

In 1971 the study enrolled a second-generation group to participate in similar examinations. It consisted of 5,124 of the original participants' adult children and their spouses. This second study is called the Framingham Offspring Study.

The main goal of this study is to identify major risk factors associated with heart disease, stroke and other diseases, and to learn the circumstances under which cardiovascular diseases arise, evolve and end fatally in the general population. The findings in this studies created a revolution in preventive medicine, and forever changed the way the medical community and general public view on the genesis of disease. In this study, there are more than 25,000 samples, each consisting of more than 100 variables. Because of the nature of this longitudinal study, some participant cannot be followed up due to their migrations. Thus, the collected data contain many missing values. During the study, cardiovascular diseases may develop for some participants, while other participants may never experience with cardiovascular diseases. This implies that some data are censored because the event of particular interest never occurred. Furthermore, data between individuals may not be independent because data for individuals in a family are clustered and likely positively correlated. Missing, censoring and clustering are common features in health studies. These three issues make data structure complicated and identification of important risk factors more challenging. In Section 6.2, we present a penalized partial likelihood approach to selecting significant risk factors for censored and clustering data. The penalized likelihood approach has been used to analyze a data subset of Framingham study in [9].

High-dimensionality is frequently seen in many other biomedical studies. For example, ecological momentary assessment data have been collected for smoking cessation studies. In such a study, each of a few hundreds participants is provided a hand-held computer, which is designed to randomly prompt the participants five to eight times per day over a period of about 50 days and to provide 50 questions at each prompt. Therefore, the data consist of a few hundreds of subjects and each of them may have more than ten thousand observed values [60]. Such data are termed intensive longitudinal data. Classical longitudinal methods are inadequate for such data. Walls and Schafer [86] presents more examples of intensive longitudinal data and some useful models to analyze this kind of data.

**2.3. Financial engineering and risk management.** Technological revolution and trade globalization have introduced a new era of financial markets. Over the last three decades, an enormous number of new financial products have been created to meet customers' demands. For example, to reduce the impact of the fluctuations of currency exchange rates on corporate finances, a multinational corporation may decide to buy options on the future of exchange rates; to reduce the risk of price fluctuations of a commodity (e.g. lumbers, corns, soybeans), a farmer may enter

into a future contract of the commodity; to reduce the risk of weather exposures, amusement parks and energy companies may decide to purchase financial derivatives based on the weather. Since the first options exchange opened in Chicago in 1973, the derivative markets have experienced extraordinary growth. Professionals in finance now routinely use sophisticated statistical techniques and modern computing power in portfolio management, securities regulation, proprietary trading, financial consulting, and risk management. For an overview, see [29] and references therein.

Complex financial markets [51] make portfolio allocation, asset pricing and risk management very challenging. For example, the price of a stock depends not only on its past values, but also its bond and derivative prices. In addition, it depends on prices of related companies and their derivatives, and on overall market conditions. Hence, the number of variables that influence asset prices can be huge and the statistical challenge is to select important factors that capture the market risks. Thanks to technological innovations, high-frequency financial data are now available for an array of different financial instruments over a long time period. The amount of financial data available to financial engineers is indeed astronomical.

Let us focus on a specific problem to illustrate the challenge of dimensionality. To optimize the performance of a portfolio [10], [12] or to manage the risk of a portfolio [70], we need to estimate the covariance matrix of the returns of assets in the portfolio. Suppose that we have 200 stocks to be selected for asset allocation. There are 20,200 parameters in the covariance matrix. This is a high-dimensional statistical problem and estimating it accurately poses challenges.

Covariance matrices pervade every facet of financial econometrics, from asset allocation, asset pricing, and risk management, to derivative pricing and proprietary trading. As mentioned earlier, they are also critical for studying genetic networks [58], as well as other statistical applications such as climatology [54]. In Section 6.1, a modified Cholesky decomposition is used to estimate huge covariance matrices using penalized least squares approach proposed in Section 2. We will introduce a factor model for covariance estimation in Section 6.3.

**2.4. Machine learning and data mining.** Machine learning and data mining extend traditional statistical techniques to handle problems with much higher dimensionality. The size of data can also be astronomical: from grocery sales and financial market trading to biomedical images and natural resource surveys. For an introduction, see the books [47], [48]. Variable selections and feature extraction are vital for such high-dimensional statistical explorations. Because of the size and complexity of the problems, the associated mathematical theory also differs from the traditional approach. The dimensionality of variables is comparable with the sample size and can even be much higher than the sample size. Selecting reliable predictors to minimize risks of prediction is fundamental to machine learning and data mining. On the other hand, as the interest mainly lies in risk minimization, unlike traditional statistics, the model parameters are only of secondary interest. As a result, crude consistency results suffice for understanding the performance of learning theory. This eases considerably

the mathematical challenges of high-dimensionality. For example, in the supervised (classification) or unsupervised (clustering) learning, we do not need to know the distributions of estimated coefficients in the underlying model. We only need to know the variables and their estimated parameters in the model. This differs from high-dimensional statistical problems in health sciences and biomedical studies, where statistical inferences are needed in presence of high-dimensionality. In Sections 4.2 and 6.4, we will address further the challenges in machine learning.

## 3. Penalized least squares

With the above background, we now consider the variable selection in the least-squares setting to gain further insights. The idea will be extended to the likelihood or pseudo-likelihood setting in the next section. We demonstrate how to directly apply the penalized least squares approach for function estimation or approximation using wavelets or spline basis, based on noisy data in Section 5. The penalized least squares method will be further extended to penalized empirical risk minimization for machine learning in Section 6.4.

Let $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \ldots, n$, be a random sample from the linear regression model

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon, \tag{3.1}$$

where $\varepsilon$ is a random error with mean 0 and finite variance $\sigma^2$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is the vector of regression coefficients. Here, we assume that all important predictors, and their interactions or functions are already in the model so that the full model (3.1) is correct.

Many variable selection criteria or procedures are closely related to minimize the following penalized least squares (PLS)

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{3.2}$$

where $d$ is the dimension of $\boldsymbol{x}$, and $p_{\lambda_j}(\cdot)$ is a penalty function, controlling model complexity. The dependence of the penalty function on $j$ allows us to incorporate prior information. For instance, we may wish to keep certain important predictors in the model and choose not to penalize their coefficients.

The form of $p_{\lambda_j}(\cdot)$ determines the general behavior of the estimator. With the entropy or $L_0$-penalty, namely, $p_{\lambda_j}(|\beta_j|) = \frac{1}{2}\lambda^2 I(|\beta_j| \neq 0)$, the PLS (3.2) becomes

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \frac{1}{2}\lambda^2 |M|, \tag{3.3}$$

where $|M| = \sum_j I(|\beta_j| \neq 0)$, the size of the candidate model. Among models with $m$ variables, the selected model is the one with the minimum residual sum of squares

(RRS), denoted by $\mathrm{RSS}_m$. A classical statistical method is to choose $m$ by maximizing the adjusted $R^2$, given by

$$R_{\mathrm{adj},m} = 1 - \frac{n-1}{n-m}\frac{\mathrm{RSS}_m}{\mathrm{RSS}_1},$$

or equivalently by minimizing $\mathrm{RSS}_m/(n-m)$, where $\mathrm{RSS}_1$ is the total sum of squares based on the null model (using the intercept only). Using $\log(1+x) \approx x$ for small $x$, it follows that

$$\log\{\mathrm{RSS}_m/(n-m)\} \approx (\log\sigma^2 - 1) + \sigma^{-2}\left\{\frac{1}{n}\mathrm{RSS}_m + \frac{1}{n}m\sigma^2\right\}. \tag{3.4}$$

Therefore, maximization of $R_{\mathrm{adj},m}$ is asymptotically equivalent to minimizing the PLS (3.3) with $\lambda = \sigma/\sqrt{n}$. Similarly, generalized cross-validation (GCV) given by

$$\mathrm{GCV}(m) = \mathrm{RSS}_m/\{n(1-m/n)^2\}$$

is asymptotically equivalent to the PLS (3.3) with $\lambda = \sqrt{2}\sigma/\sqrt{n}$ and so is the cross-validation (CV) criterion.

Many popular variable selection criteria can be shown asymptotically equivalent to the PLS (3.3) with appropriate values of $\lambda$, though these criteria were motivated from different principles. See [69] and references therein. For instance, RIC [38] corresponds to $\lambda = \sqrt{2\log(d)}(\sigma/\sqrt{n})$. Since the entropy penalty function is discontinuous, minimizing the entropy-penalized least-squares requires exhaustive search, which is not feasible for high-dimensional problem. In addition, the sampling distributions of resulting estimates are hard to derive.

Many researchers have been working on minimizing the PLS (3.2) with $L_p$-penalty for some $p > 0$. It is well known that the $L_2$-penalty results in a ridge regression estimator, which regularizes and stabilizes the estimator but introduces biases. However, it does not shrink any coefficients directly to zero.

The $L_p$-penalty with $0 < p < 2$ yields bridge regression [39], intermediating the best-subset ($L_0$-penalty) and the ridge regression ($L_2$-penalty). The non-negative garrote [8] shares the same spirit as that of bridge regression. With the $L_1$-penalty specifically, the PLS estimator is called LASSO in [81]. In a seminal paper, Donoho and Elad [20] show that penalized $L_0$-solution can be found by using penalized $L_1$-method for sparse problem. When $p \leq 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients.

Antoniadis and Fan [1] discussed how to choose a penalty function for wavelets regression. Fan and Li [33] advocated penalty functions with three properties:

**a.** *Sparsity:* The resulting estimator should automatically set small estimated coefficients to zero to accomplish variable selection.

**b.** *Unbiasedness:* The resulting estimator should have low bias, especially when the true coefficient $\beta_j$ is large.

**c.** *Continuity:* The resulting estimator should be continuous to reduce instability in model prediction.

To gain further insights, let us assume that the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ for model (3.1) is orthogonal and satisfies that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = I_d$. Let $\boldsymbol{z} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ be the least squares estimate of $\boldsymbol{\beta}$. Then (3.2) becomes

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z}\| + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

Thus the PLS reduces to a componentwise minimization problem:

$$\min_{\beta_j}\left\{\frac{1}{2}(z_j - \beta_j)^2 + p_{\lambda_j}(|\beta_j|)\right\}, \quad \text{for } j = 1, \ldots, d,$$

where $z_j$ is the $j$-th component of $\boldsymbol{z}$. Suppress the subscript $j$ and let

$$Q(\beta) = \frac{1}{2}(z - \beta)^2 + p_\lambda(|\beta|). \tag{3.5}$$

Then the first order derivative of $Q(\beta)$ is given by

$$Q'(\beta) = \beta - z + p'_\lambda(|\beta|)\mathrm{sgn}(\beta) = \mathrm{sgn}(\beta)\{|\beta| + p'_\lambda(|\beta|)\} - z.$$

Antoniadis and Fan [1] and Fan and Li [33] derived that the PLS estimator possesses the following properties:

(a) *sparsity* if $\min_\beta\{|\beta| + p'_\lambda(|\beta|)\} > 0$;

(b) *unbiasedness* $p'_\lambda(|\beta|) = 0$ for large $|\beta|$;

(c) *continuity* if and only if $\mathrm{argmin}_\beta\{|\beta| + p'_\lambda(|\beta|)\} = 0$.

The $L_p$-penalty with $0 \le p < 1$ does not satisfy the continuity condition, the $L_1$ penalty does not satisfy the unbiasedness condition, and $L_p$ with $p > 1$ does not satisfy the sparsity condition. Therefore, none of the $L_p$-penalties satisfies the above three conditions simultaneously, and $L_1$-penalty is the such penalty that is both convex and produces sparse solutions. Of course, the class of penalty functions satisfying the aforementioned three conditions are infinitely many. Fan and Li [33] suggested the use of the smoothly clipped absolute deviation (SCAD) penalty defined as

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \le |\beta| < \lambda; \\ -(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)/\{2(a-1)\}, & \text{if } \lambda \le |\beta| < a\lambda; \\ (a+1)\lambda^2/2, & \text{if } |\beta| \ge a\lambda. \end{cases}$$

They further suggested using $a = 3.7$. This function has similar feature to the penalty function $\lambda|\beta|/(1 + |\beta|)$ advocated in [71]. Figure 1 depicts the SCAD, $L_{0.5}$-penalty,

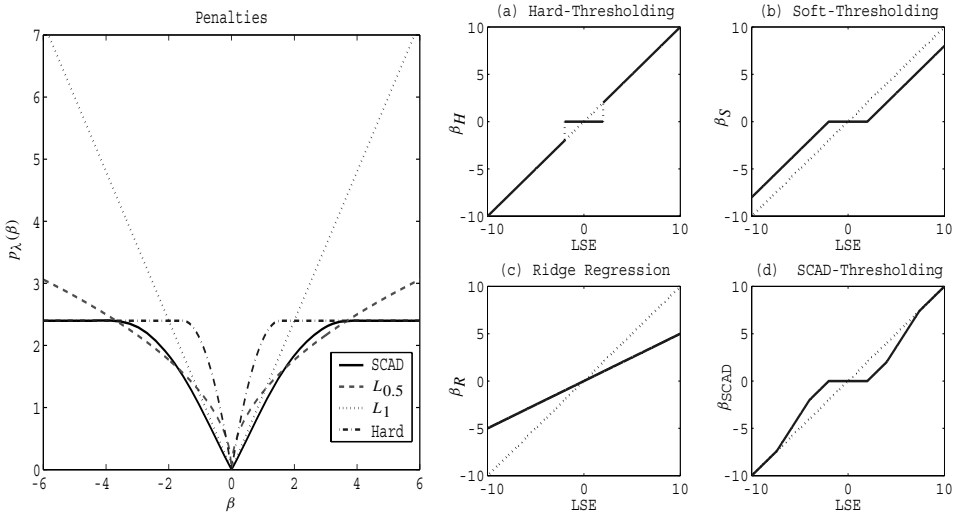



Figure 1. Penalty functions (left panel) and PLS estimators (right panel).

$L_1$-penalty, and hard thresholding penalty (to be introduced) functions. These four penalty functions are singular at the origin, a necessary condition for sparsity in variable selection. Furthermore, the SCAD, hard-thresholding and $L_{0.5}$ penalties are nonconvex over $(0, +\infty)$ in order to reduce the estimation bias.

Minimizing the PLS (3.5) with the entropy penalty or hard-thresholding penalty $p_\lambda(\beta) = \lambda^2 - (\lambda - |\beta|)_+^2$ (which is smoother) yields the hard-thresholding rule [23] $\hat{\beta}_H = zI(|z| > \lambda)$. With the $L_1$-penalty, the PLS estimator is $\hat{\beta}_S = \text{sgn}(z)(|z| - \lambda)_+$, the soft-thresholding rule [3], [23]. The $L_2$-penalty results in the ridge regression $\hat{\beta}_R = (1 + \lambda)^{-1} z$ and the SCAD penalty gives the solution

$$\hat{\beta}_{\text{SCAD}} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \le 2\lambda; \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \le a\lambda; \\ z, & \text{when } |z| > a\lambda. \end{cases}$$

These functions are also shown in Figure 1. The SCAD is an improvement over the $L_0$-penalty in two aspects: saving computational cost and resulting in a continuous solution to avoid unnecessary modeling variation. Furthermore, the SCAD improves bridge regression by reducing modeling variation in model prediction. Although similar in spirit to the $L_1$-penalty, the SCAD also improves the $L_1$-penalty by avoiding excessive estimation bias since the solution of the $L_1$-penalty could shrink all regression coefficients by a constant, e.g., the soft thresholding rule.

## 4. Penalized likelihood

PLS can easily be extended to handle a variety of response variables, including binary response, counts, and continuous response. A popular family of this kind is called generalized linear models. Our approach can also be applied to the case where the likelihood is a quasi-likelihood or other discrepancy functions. This will be demonstrated in Section 6.2 for analysis of survival data, and in Section 6.4 for machine learning.

Suppose that conditioning on $\boldsymbol{x}_i$, $y_i$ has a density $f\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\}$, where $g$ is a known inverse link function. Define a penalized likelihood as

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log f\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\} - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{4.1}$$

Maximizing the penalized likelihood results in a penalized likelihood estimator. For certain penalties, such as the SCAD, the selected model based on the nonconcave penalized likelihood satisfies $\beta_j = 0$ for certain $\beta_j$'s. Therefore, parameter estimation is performed at the same time as the model selection.

**Example** (*Logistics Regression*). Suppose that given $\boldsymbol{x}_i$, $y_i$ follows a Bernoulli distribution with success probability $P\{y_i = 1|\boldsymbol{x}_i\} = p(\boldsymbol{x}_i)$. Take $g(u) = \exp(u)/(1 + \exp(u))$, i.e. $p(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})\}$. Then (4.1) becomes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\}] - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

Thus, variable selection for logistics regression can be achieved by maximizing the above penalized likelihood.

**Example** (*Poisson Log-linear Regression*). Suppose that given $\boldsymbol{x}_i$, $y_i$ follows a Poisson distribution with mean $\lambda(\boldsymbol{x}_i)$. Take $g(\cdot)$ to be the log-link, i.e. $\lambda(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Then (4.1) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\} - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|)$$

after dropping a constant. Thus, maximizing the above penalized likelihood with certain penalty functions yields a sparse solution for $\boldsymbol{\beta}$.

**4.1. Oracle properties.** Maximizing a penalized likelihood selects variables and estimates parameters simultaneously. This allows us to establish the sampling properties of the resulting estimators. Under certain regularity conditions, Fan and Li [33] demonstrated how the rates of convergence for the penalized likelihood estimators

depend on the regularization parameter $\lambda_n$ and established the oracle properties of the penalized likelihood estimators.

In the context of variable selection for high-dimensional modeling, it is natural to allow the number of introduced variables to grow with the sample sizes. Fan and Peng [35] have studied the asymptotic properties of the penalized likelihood estimator for situations in which the number of parameters, denoted by $d_n$, tends to $\infty$ as the sample size $n$ increases. Denote $\boldsymbol{\beta}_{n0}$ to be the true value of $\boldsymbol{\beta}$. To emphasize the dependence of $\lambda_j$ on $n$, we use notation $\lambda_{n,j}$ for $\lambda_j$ in this subsection. Define

$$a_n = \max\{p'_{\lambda_{n,j}}(|\beta_{n0j}|) : \beta_{n0j} \neq 0\} \quad \text{and} \quad b_n = \max\{|p''_{\lambda_{n,j}}(|\beta_{n0j}|)| : \beta_{n0j} \neq 0\}. \tag{4.2}$$

Fan and Peng [35] showed that if both $a_n$ and $b_n$ tend to 0 as $n \to \infty$, then under certain regularity conditions, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{n0}\| = O_P\{\sqrt{d_n}(n^{-1/2} + a_n)\}. \tag{4.3}$$

It is clear from (4.3) that by choosing a proper $\lambda_{n,j}$ such that $a_n = O(n^{-1/2})$, there exists a root-$(n/d_n)$ consistent penalized likelihood estimator. For example, for the SCAD, the penalized likelihood estimator is root-$(n/d_n)$ consistent if all $\lambda_{n,j}$'s tend to 0.

Without loss of generality assume that, unknown to us, the first $s_n$ components of $\boldsymbol{\beta}_{n0}$, denoted by $\boldsymbol{\beta}_{n01}$, are nonzero and do not vanish and the remaining $d_n - s_n$ coefficients, denoted by $\boldsymbol{\beta}_{n02}$, are 0. Denote by

$$\Sigma = \operatorname{diag}\bigl\{p''_{\lambda_{n,1}}(|\beta_{n01}|), \ldots, p''_{\lambda_{n,s_n}}(|\beta_{n0s_n}|)\bigr\}$$

and

$$\boldsymbol{b} = \bigl(p'_{\lambda_{n,1}}(|\beta_{n01}|)\operatorname{sgn}(\beta_{n01}), \ldots, p'_{\lambda_{n,s_n}}(|\beta_{n0s_n}|)\operatorname{sgn}(\beta_{n0s_n})\bigr)^T.$$

**Theorem 1.** *Assume that as* $n \to \infty$, $\min_{1\le j\le s_n} |\beta_{n0j}|/\lambda_{n,j} \to \infty$ *and that the penalty function* $p_{\lambda_j}(|\beta_j|)$ *satisfies*

$$\liminf_{n\to\infty} \liminf_{\beta_j \to 0+} p'_{\lambda_{n,j}}(\beta_j)/\lambda_{n,j} > 0. \tag{4.4}$$

*If* $\lambda_{n,j} \to 0$, $\sqrt{n/d_n}\lambda_{n,j} \to \infty$ *and* $d_n^5/n \to 0$ *as* $n \to \infty$, *then with probability tending to* 1, *the root* $n/d_n$ *consistent local maximizers* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T)^T$ *must satisfy:*

(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$;

(ii) *Asymptotic normality: for any* $q \times s_n$ *matrix* $\boldsymbol{A}_n$ *such that* $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \boldsymbol{G}$, *a* $q \times q$ *positive definite symmetric matrix,*

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_1^{-1/2}\{\boldsymbol{I}_1 + \Sigma\}\bigl\{\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10} + (\boldsymbol{I}_1 + \Sigma)^{-1}\boldsymbol{b}\bigr\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{G})$$

*where* $\boldsymbol{I}_1 = \boldsymbol{I}_1(\boldsymbol{\beta}_{n10}, \mathbf{0})$, *the Fisher information knowing* $\boldsymbol{\beta}_{n20} = \mathbf{0}$.

The theorem implies that any finite set of elements of $\hat{\boldsymbol{\beta}}_{n1}$ are jointly asymptotically normal. For the SCAD, if all $\lambda_{j,n} \to 0$, $a_n = 0$. Hence, when $\sqrt{n/d_n}\lambda_{n,j} \to \infty$, its corresponding penalized likelihood estimators possess the oracle property, i.e., perform as well as the maximum likelihood estimates for estimating $\boldsymbol{\beta}_{n1}$ knowing $\boldsymbol{\beta}_{n2} = \mathbf{0}$. That is, with probability approaching to 1,

$$\hat{\boldsymbol{\beta}}_{n2} = 0, \quad \text{and} \quad \sqrt{n}\boldsymbol{A}_n \boldsymbol{I}_1^{1/2}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10}) \to N(\mathbf{0}, \boldsymbol{G}).$$

For the $L_1$-penalty, $a_n = \max_j \lambda_{j,n}$. Hence, the root-$n/d_n$ consistency requires that $\lambda_{n,j} = O(\sqrt{d_n/n})$. On the other hand, the oracle property in Theorem 2 requires that $\sqrt{n/d_n}\lambda_{n,j} \to \infty$. These two conditions for LASSO cannot be satisfied simultaneously. It has indeed been shown that the oracle property does not hold for the $L_1$-penalty even in the finite parameter setting [90].

**4.2. Risk minimization and persistence.** In machine learning such as tumor classifications, the primary interest centers on the misclassification errors or more generally expected losses, not the accuracy of estimated parameters. This kind of properties is called persistence in [42], [43].

Consider predicting the response $Y$ using a class of model $g(\boldsymbol{x}^T\boldsymbol{\beta})$ with a loss function $\ell\{g(\boldsymbol{X}^T\boldsymbol{\beta}), Y)$. Then the risk is

$$L_n(\boldsymbol{\beta}) = E\ell\{g(\boldsymbol{X}^T\boldsymbol{\beta}), Y\},$$

where $n$ is used to stress the dependence of dimensionality $d$ on $n$. The minimum risk is obtained at $\boldsymbol{\beta}_n^* = \operatorname{argmin}_{\boldsymbol{\beta}} L_n(\boldsymbol{\beta})$. In the likelihood context, $\ell = -\log f$. Suppose that there is an estimator $\hat{\boldsymbol{\beta}}_n$ based on a sample of size $n$. This can be done by the penalized empirical risk minimization similarly to (4.1):

$$n^{-1}\sum_{i=1}^{n} \ell\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\} + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{4.5}$$

based on a set of training data $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$. The persistence requires

$$L_n(\hat{\boldsymbol{\beta}}_n) - L_n(\boldsymbol{\beta}_n^*) \xrightarrow{P} 0, \tag{4.6}$$

but not necessarily the consistency of $\hat{\boldsymbol{\beta}}_n$ to $\boldsymbol{\beta}_n^*$. This is in general a much weaker mathematical requirement. Greenshtein and Ritov [43] show that if the non-sparsity rate $s_n = O\{(n/\log n)^{1/2}\}$ and $d_n = n^\alpha$ for some $\alpha > 1$, LASSO (penalized $L_1$ least-squares) is persistent under the quadratic loss. Greenshtein [42] extends the results to the case where $s_n = O\{n/\log n\}$ and more general loss functions. Meinshausen [66] considers a case with finite non-sparsity $s_n$ but with $\log d_n = n^\xi$, with $\xi \in (0, 1)$. It is shown there that for the quadratic loss, LASSO is persistent, but the rate to persistency is slower than a relaxed LASSO. This again shows the bias problems in LASSO.

**4.3. Issues in practical implementation.** In this section, we address practical implementation issues related to the PLS and penalized likelihood.

**Local quadratic approximation (LQA).** The $L_p$, $(0 < p < 1)$, and SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. Therefore, maximizing the nonconcave penalized likelihood is challenging. Fan and Li [33] propose locally approximating them by a quadratic function as follows. Suppose that we are given an initial value $\boldsymbol{\beta}^0$ that is close to the optimizer of $Q(\boldsymbol{\beta})$. For example, take initial value to be the maximum likelihood estimate (without penalty). Under some regularity conditions, the initial value is a consistent estimate for $\boldsymbol{\beta}$, and therefore it is close to the true value. Thus, we can locally approximate the penalty function by a quadratic function as

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\beta_j^0|) + \frac{1}{2}\{p'_{\lambda_n}(|\beta_j^0|)/|\beta_j^0|\}(\beta_j^2 - \beta_j^{02}), \quad \text{for } \beta_j \approx \beta_j^0. \tag{4.7}$$

To avoid numerical instability, we set $\hat{\beta}_j = 0$ if $\beta_j^0$ is very close to 0. This corresponds to deleting $x_j$ from the final model. With the aid of the LQA, the optimization of penalized least-squares, penalized likelihood or penalized partial likelihood (see Section 6.2) can be carried out by using the Newton–Raphson algorithm. It is worth noting that the LQA should be updated at each step during the course of iteration of the algorithm. We refer to the modified Newton–Raphson algorithm as the LQA algorithm.

The convergence property of the LQA algorithm was studied in [52], whose authors first showed that the LQA plays the same role as the E-step in the EM algorithm [18]. Therefore the behavior of the LQA algorithm is similar to the EM algorithm. Unlike the original EM algorithm, in which a full iteration for maximization is carried out after every E-step, we update the LQA at each step during the iteration course. This speeds up the convergence of the algorithm. The convergence rate of the LQA algorithm is quadratic which is the same as that of the modified EM algorithm [56].

When the algorithm converges, the estimator satisfies the condition

$$\partial \ell(\hat{\boldsymbol{\beta}})/\partial \beta_j + n p'_{\lambda_j}(|\hat{\beta}_j|)\operatorname{sgn}(\hat{\beta}_j) = 0,$$

the penalized likelihood equation, for non-zero elements of $\hat{\boldsymbol{\beta}}$.

**Standard error formula.** Following conventional techniques in the likelihood setting, we can estimate the standard error of the resulting estimator by using the sandwich formula. Specifically, the corresponding sandwich formula can be used as an estimator for the covariance of the estimator $\hat{\boldsymbol{\beta}}_1$, the non-vanishing component of $\hat{\boldsymbol{\beta}}$. That is,

$$\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}_1) = \{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1} \widehat{\operatorname{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}}_1)\}\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1}, \tag{4.8}$$

where $\widehat{\operatorname{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}}_1)\}$ is the usual empirically estimated covariance matrix and

$$\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1) = \operatorname{diag}\{p'_{\lambda_1}(|\hat{\beta}_1|)/|\hat{\beta}_1|, \ldots, p'_{\lambda_{s_n}}(|\hat{\beta}_{s_n}|)/|\hat{\beta}_{s_n}|\}$$

and $s_n$ the dimension of $\hat{\boldsymbol{\beta}}_1$. Fan and Peng [35] demonstrated the consistency of the sandwich formula:

**Theorem 2.** *Under the conditions of Theorem* 1*, we have*

$$\boldsymbol{A}_n \widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_1)\boldsymbol{A}_n^T - \boldsymbol{A}_n \Sigma_n \boldsymbol{A}_n^T \xrightarrow{P} 0 \quad \text{as } n \to \infty$$

*for any matrix* $\boldsymbol{A}_n$ *such that* $\boldsymbol{A}_n \boldsymbol{A}_n^T = \boldsymbol{G}$*, where* $\Sigma_n = (\boldsymbol{I}_1 + \Sigma)^{-1}\boldsymbol{I}_1^{-1}(\boldsymbol{I}_1 + \Sigma)^{-1}$.

**Selection of regularization parameters.** To implement the methods described in previous sections, it is desirable to have an automatic method for selecting the thresholding parameter $\lambda$ in $p_\lambda(\cdot)$ based on data. Here, we estimate $\lambda$ via minimizing an approximate generalized cross-validation (GCV) statistic in [11]. By some straightforward calculation, the effective number of parameters for $Q(\boldsymbol{\beta})$ in the last step of the Newton–Raphson algorithm iteration is

$$e(\boldsymbol{\lambda}) \equiv e(\lambda_1, \ldots, \lambda_d) = \mathrm{tr}[\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\nabla^2 \ell(\hat{\boldsymbol{\beta}})].$$

Therefore the generalized cross-validation statistic is defined by

$$\mathrm{GCV}(\boldsymbol{\lambda}) = -\ell(\hat{\boldsymbol{\beta}})/[n\{1 - e(\boldsymbol{\lambda})/n\}^2]$$

and $\hat{\boldsymbol{\lambda}} = \mathrm{argmin}_{\boldsymbol{\lambda}}\{\mathrm{GCV}(\boldsymbol{\lambda})\}$ is selected.

To find an optimal $\boldsymbol{\lambda}$, we need to minimize the GCV over a $d_n$-dimensional space. This is an unduly onerous task. Intuitively, it is expected that the magnitude of $\lambda_j$ should be proportional to the standard error of the maximum likelihood estimate of $\beta_j$. Thus, we set $\boldsymbol{\lambda} = \lambda \mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ in practice, where $\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ denotes the standard error of the MLE. Therefore, we minimize the GCV score over the one-dimensional space, which will save a great deal of computational cost. The behavior of such a method has been investigated recently.

## 5. Applications to function estimation

Let us begin with one-dimensional function estimation. Suppose that we have noisy data at possibly irregular design points $\{x_1, \ldots, x_n\}$:

$$y_i = m(x_i) + \varepsilon_i,$$

where $m$ is an unknown regression and $\varepsilon_i$'s are iid random error following $N(0, \sigma^2)$. Local modeling techniques [31] have been widely used to estimate $m(\cdot)$. Here we focus on global function approximation methods.

Wavelet transforms are a device for representing functions in a way that is local in both time and frequency domains [13], [14], [64], [65]. During the last decade, they have received a great deal of attention in applied mathematics, image analysis, signal

compression, and many other fields of engineering. Daubechies [17] and Meyer [68] are good introductory references to this subject. Wavelet-based methods have many exciting statistical properties [23]. Earlier papers on wavelets assume the regular design points, i.e, $x_i = \frac{i}{n}$ (usually $n = 2^k$ for some integer $k$) so that fast computation algorithms can be implemented. See [24] and references therein. For an overview of wavelets in statistics, see [87].

Antoniadis and Fan [1] discussed how to apply wavelet methods for function estimation with irregular design points using penalized least squares. Without loss of generality, assume that $m(x)$ is defined on [0, 1]. By moving nondyadic points to dyadic points, we assume $x_i = n_i/2^J$ for some $n_i$ and some fine resolution $J$ that is determined by users. To make this approximation errors negligible, we take $J$ large enough such that $2^J \geq n$. Let $\boldsymbol{W}$ be a given wavelet transform at all dyadic points $\{i/2^J :, i = 1, \ldots, 2^J - 1\}$. Let $N = 2^J$ and $\boldsymbol{a}_i$ be the $n_i$-th column of $\boldsymbol{W}$, an $N \times N$ matrix, and $\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{m}$ be the wavelet transform of the function $m$ at dyadic points. Then it is easy to see that $m(x_i) = \boldsymbol{a}_i^T \boldsymbol{\beta}$. This yields an overparameterized linear model

$$y_i = \boldsymbol{a}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{5.1}$$

which aims at reducing modeling biases. However, one cannot find a reasonable estimate of $\boldsymbol{\beta}$ by using the ordinary least squares method since $N \geq n$. Directly applying penalized least squares, we have

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{a}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{N} p_{\lambda_j}(|\beta_j|). \tag{5.2}$$

If the sampling points are equally spaced and $n = 2^J$, the corresponding design matrix of linear model (5.1) becomes a square orthogonal matrix. From the discussion in Section 3, minimizing the PLS (5.2) with the entropy penalty or the hard-thresholding penalty results in a hard-thresholding rule. With the $L_1$ penalty, the PLS estimator is the soft-thresholding rule. Assume that $p_\lambda(\cdot)$ is nonnegative, nondecreasing, and differentiable over $(0, \infty)$ and that function $-\beta - p'_\lambda(\beta)$ is strictly unimodal on $(0, \infty)$, $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$. Then Antoniadis and Fan [1] showed that the resulting penalized least-squares estimator that minimizes (5.2) is adaptively minimax within a factor of logarithmic order as follows. Define the Besov space ball $B^r_{p,q}(C)$ to be

$$B^r_{p,q}(C) = \left\{ m \in L_p : \textstyle\sum_j (2^{j(r+1/2-1/p)} \|\boldsymbol{\theta}_{j\cdot}\|_p)^q < C \right\},$$

where $\boldsymbol{\theta}_{j\cdot}$ is the vector of wavelet coefficients of function $m$ at the resolution level $j$. Here $r$ indicates the degree of smoothness of the regression functions $m$.

**Theorem 3.** *Suppose that the regression function $m(\cdot)$ is in a Besov ball with $r + 1/2 - 1/p > 0$. Then the maximum risk of the PLS estimator $\hat{m}(\cdot)$ over $B^r_{p,q}(C)$ is of rate $O(n^{-2r/(2r+1)} \log(n))$ when the universal thresholding $\sqrt{2\log(n)/n}$ is used. It also achieves the rate of convergence $O\{n^{-2r/(2r+1)} \log(n)\}$ when the minimax thresholding $p_n/\sqrt{n}$ is used, where $p_n$ is given in* [1].

We next consider multivariate regression function estimation. Suppose that $\{\boldsymbol{x}_i, y_i\}$ is a random sample from the regression model

$$y = m(\boldsymbol{x}) + \varepsilon,$$

where, without loss of generality, it is assumed that $\boldsymbol{x} \in [0, 1]^d$. Radial basis and neural-network are also popular for approximating multi-dimensional functions. In the literature of spline smoothing, it is typically assumed that the mean function $m(\boldsymbol{x})$ has a low-dimensional structure. For example,

$$m(\boldsymbol{x}) = \mu_0 + \sum_j m_j(x_j) + \sum_{k<l} m_{kl}(x_k, x_l).$$

For given knots, a set of spline basis functions can be constructed. The two most popular spline bases are the truncated power spline basis $1, x, x^2, x^3, (x - t_j)_+^3$, $(j = 1, \ldots, J)$, where $t_j$'s are knots, and the B-spline basis (see [6] for definition). The $B$-spline basis is numerically more stable since the multiple correlation among the basis functions is smaller, but the power truncated spline basis has the advantage that deleting a basis function is the same as deleting a knot.

For a given set of 1-dimensional spline bases, we can further construct a multivariate spline basis using tensor products. Let $\{B_1, \ldots, B_J\}$ be a set of spline basis functions on $[0, 1]^d$. Approximate the regression function $m(\boldsymbol{x})$ by a linear combination of the basis functions, $\sum \beta_j B_j(\boldsymbol{x})$, say. To avoid a large approximation bias, we take a large $J$. This yields an overparameterized linear model, and the fitted curve of the least squares estimate is typically undersmooth. Smoothing spline suggested penalizing the roughness of the resulting estimate. This is equivalent to the penalized least squares with a quadratic penalty. In a series of work by Stone and his collaborators (see [76]), they advocate using regression splines and modifying traditional variable selection approaches to select useful spline subbases. Ruppert *et al.* [75] advocated penalized splines in statistical modeling, in which power truncated splines are used with the $L_2$ penalty. Another kind of penalized splines method proposed by [28] shares the same spirit of [75].

## 6. Some solutions to the challenges

In this section, we provide some solutions to problems raised in Section 2.

**6.1. Computational biology.** As discussed in Section 2.1, the first statistical challenge in computational biology is how to remove systematic biases due to experiment variations. Thus, let us first discuss the issue of normalization of cDNA-microarrays. Let $Y_g$ be the log-ratio of the intensity of gene $g$ of the treatment sample over that of the control sample. Denote by $X_g$ the average of the log-intensities of gene $g$ at the treatment and control samples. Set $r_g$ and $c_g$ be the row and column of the block

where the cDNA of gene $g$ resides. Fan *et al.* [37] use the following model to estimate the intensity and block effect:

$$Y_g = \alpha_g + \beta_{r_g} + \gamma_{c_g} + f(X_g) + \varepsilon_g, \quad g = 1, \ldots, N \tag{6.1}$$

where $\alpha_g$ is the treatment effect on gene $g$, $\beta_{r_g}$ and $\gamma_{c_g}$ are block effects that are decomposed into the column and row effect, $f(X_g)$ represents the intensity effect and $N$ is the total number of genes. Based on $J$ arrays, an aim of microarray data analysis is to find genes $g$ with $\alpha_g$ statistically significantly different from 0. However, before carrying multiple array comparisons, the block and treatment effects should first be estimated and removed. For this normalization purpose, parameters $\alpha_g$ are nuisance and high-dimensional (recall $N$ is in the order of tens of thousands). On the other hand, the number of significantly expressed genes is relatively small, yielding the sparsity structure of $\alpha_g$.

Model (6.1) is not identifiable. Fan *et al.* [37] use within-array replicates to infer about the block and treatment effects. Suppose that we have $I$ replications for $G$ genes, which could be a small fraction of $N$. For example, in [37], only 111 genes were repeated at random blocks ($G = 111$, $I = 2$), whereas in [63], all genes were repeated three times, i.e. $I = 3$ and $N = 3G$, though both have about $N \approx 20,000$ genes printed on an array. Using $I$ replicated data on $G$ genes, model (6.1) becomes

$$Y_{gi} = \alpha_g + \beta_{r_{gi}} + \gamma_{c_{gi}} + f(X_{gi}) + \varepsilon_{gi}, \quad g = 1, \ldots, G, \ i = 1, \ldots, I. \tag{6.2}$$

With estimated coefficients $\hat{\beta}$ and $\hat{\gamma}$ and the function $\hat{f}$, model (6.1) implies that the normalized data are $Y_g^* = Y_g - \hat{\beta}_{r_g} - \hat{\gamma}_{c_g} - \hat{f}(X_g)$ even for non-repeated genes.

Model (6.2) can be used to remove the intensity effect array by array, though the number of nuisance parameters is very large, a fraction of total sample size in (6.2). To improve the efficiency of estimation, Fan *et al.* [36] aggregate the information from other microarrays (total $J$ arrays):

$$Y_{gij} = \alpha_g + \beta_{r_{gi},j} + \gamma_{c_{gi},j} + f_j(X_{gij}) + \varepsilon_{gi}, \quad j = 1, \ldots, J, \tag{6.3}$$

where the subscript $j$ denotes the array effect.

The parameters in (6.2) can be estimated by the profile least-squares method using the Gauss–Seidel type of algorithm. See [36] for details. To state the results, let us write model (6.2) as

$$Y_{gi} = \alpha_g + \boldsymbol{Z}_{gi}^T \boldsymbol{\beta} + f(X_{gi}) + \varepsilon_{gi}, \tag{6.4}$$

by appropriately introducing the dummy variable $\boldsymbol{Z}$. Fan *et al.* [36] obtained the following results.

**Theorem 4.** *Under some regularity conditions, as $n = IG \to \infty$, the profile least-squares estimator of model* (6.4) *has*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N\Big(0, \frac{I}{I-1}\sigma^2 \Sigma^{-1}\Big),$$

*where* $\Sigma = E\{\mathrm{Var}(\mathbf{Z}|X)\}$ *and* $\sigma^2 = \mathrm{Var}(\varepsilon)$. *In addition,* $\hat{f}(x) - f(x) = O_P(n^{-2/5})$.

**Theorem 5.** *Under some regularity conditions, as* $n = IG \to \infty$, *when* $X$ *and* $\mathbf{Z}$ *are independent, the profile least-squares estimator based on* (6.3) *possesses*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \xrightarrow{\mathcal{D}} N\Big(0, \frac{I(J-1)+1}{J(I-1)}\,\sigma^2\Sigma^{-1}\Big).$$

The above theorems show that the block effect can be estimated at rate $O_P(n^{-1/2})$ and intensity effect $f$ can be estimated at rate $O_P(n^{-2/5})$. This rate can be improved to $O_P(n^{-1/2} + N^{-2/5})$ when data in (6.1) are all used. The techniques have also been adapted for the normalization of Affymetrix arrays [30]. Once the arrays have been normalized, the problem becomes selecting significantly expressed genes using the normalized data

$$Y^*_{gj} = \alpha_g + \varepsilon_{gj}, \quad g = 1, \ldots, N,\ j = 1, \ldots, J, \tag{6.5}$$

where $Y^*_{gj}$ is the normalized expression of gene $g$ in array $j$. This is again a high-dimensional statistical inference problem. The issues of computing P-values and false discovery are given in Section 2.1.

Estimation of high-dimensional covariance matrices is critical in studying genetic networks. PLS and penalized likelihood can be used to estimate large scale covariance matrices effectively and parsimoniously [49], [59]. Let $\boldsymbol{w} = (W_1, \ldots, W_d)^T$ be a $d$-dimensional random vector with mean zero and covariance $\Sigma$. Using the modified Cholesky decomposition, we have $\boldsymbol{L}\Sigma\boldsymbol{L}^T = \boldsymbol{D}$, where $\boldsymbol{L}$ is a lower triangular matrix having ones on its diagonal and typical element $-\phi_{tj}$ in the $(t, j)$th position for $1 \le j < t \le d$, and $\boldsymbol{D} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_d^2)^T$ is a diagonal matrix. Denote $\boldsymbol{e} = \boldsymbol{L}\boldsymbol{w} = (e_1, \ldots, e_d)^T$. Since $\boldsymbol{D}$ is diagonal, $e_1, \ldots, e_d$ are uncorrelated. Thus, for $2 \le t \le d$

$$W_t = \sum_{j=1}^{t-1} \phi_{tj} W_j + e_t. \tag{6.6}$$

That is, the $W_t$ is an autoregressive (AR) series, which gives an interpretation for elements of $L$ and $D$, and allows us to use PLS for covariance selection. We first estimate $\sigma_t^2$ using the mean squared errors of model (6.6). Suppose that $\boldsymbol{w}_i$, $i = 1, \ldots, n$, is a random sample from $\boldsymbol{w}$. For $t = 2, \ldots, d$, covariance selection can be achieved by minimizing the following PLS functions:

$$\frac{1}{2n}\sum_{i=1}^{n}\Big(W_{it} - \sum_{j=1}^{t-1}\phi_{tj}W_{ij}\Big)^2 + \sum_{j=1}^{t-1} p_{\lambda_{t,j}}(|\phi_{tj}|). \tag{6.7}$$

This reduces the non-sparse elements in the lower triangle matrix $\boldsymbol{L}$. With estimated $\boldsymbol{L}$, the diagonal elements can be estimated by the sample variance of the components in $\hat{L}\boldsymbol{w}_i$. The approach can easily be adapted to estimate the sparse precision matrix $\Sigma^{-1}$. See [67] for a similar approach and a thorough study.

**6.2. Health studies.** Survival data analysis has been a very active research topic because survival data are frequently collected from reliability analysis, medical studies, and credit risks. In practice, many covariates are often available as potential risk factors. Selecting significant variables plays a crucial role in model building for survival data but is challenging due to the complicated data structure. Fan and Li [34] derived the nonconcave penalized partial likelihood for Cox's model and Cox's frailty model, the most commonly used semiparametric models in survival analysis. Cai *et al.* [9] proposed a penalized pseudo partial likelihood for marginal Cox's model with multivariate survival data and applied the proposed methodology for a subset data in the Framingham study, introduced in Section 2.2.

Let $T$, $C$ and $\boldsymbol{x}$ be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It is assumed that $T$ and $C$ are conditionally independent given $\boldsymbol{x}$, that the censoring mechanism is noninformative, and that the observed data $\{(\boldsymbol{x}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ is an independently and identically distributed random sample from a certain population $(\boldsymbol{x}, Z, \delta)$. The Cox model assumes the conditional hazard function of $T$ given $\boldsymbol{x}$

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{6.8}$$

where $h_0(t)$ is an unspecified baseline hazard function. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let $(j)$ provide the label for the item failing at $t_j^0$ so that the covariates associated with the $N$ failures are $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(N)}$. Let $R_j$ denote the risk set right before the time $t_j^0 : R_j = \{i : Z_i \geq t_j^0\}$. Fan and Li [34] proposed the penalized partial likelihood

$$Q(\boldsymbol{\beta}) = \sum_{j=1}^{N}\Big[\boldsymbol{x}_{(j)}^T\boldsymbol{\beta} - \log\Big\{\sum_{i\in R_j}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\Big\}\Big] - n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{6.9}$$

The penalized likelihood estimate of $\boldsymbol{\beta}$ is to maximize (6.9) with respect to $\boldsymbol{\beta}$.

For finite parameter settings, Fan and Li [34] showed that under certain regularity conditions, if both $a_n$ and $b_n$ tend to 0, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of the penalized partial likelihood function in (6.9) such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$. They further demonstrated the following oracle property.

**Theorem 6.** *Assume that the penalty function $p_{\lambda_n}(|\beta|)$ satisfies condition* (4.4). *If $\lambda_{n,j} \to 0$, $\sqrt{n}\lambda_{n,j} \to \infty$ and $a_n = O(n^{-1/2})$, then under some mild regularity conditions, with probability tending to* 1, *the root n consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ of $Q(\boldsymbol{\beta})$ defined in* (6.9) *must satisfy*

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}, \quad \textit{and} \quad \sqrt{n}(I_1 + \Sigma)\big\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1 + \Sigma)^{-1}\boldsymbol{b}\big\} \xrightarrow{\mathcal{D}} N\big\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\big\},$$

*where $I_1$ is the first $s \times s$ submatrix of the Fisher information matrix $I(\boldsymbol{\beta}_0)$ of the partial likelihood.*

Cai *et al.* [9] investigated the sampling properties of penalized partial likelihood estimate with a diverging number of predictors and clustered survival data. They showed that the oracle property is still valid for penalized partial likelihood estimation for the Cox marginal models with multivariate survival data.

**6.3. Financial engineering and risk management.** There are many outstanding challenges of dimensionality in diverse fields of financial engineering and risk management. To be concise, we focus only on the issue of covariance matrix estimation using a factor model.

Let $Y_i$ be the excess return of the $i$-th asset over the risk-free asset. Let $f_1, \ldots, f_K$ be the factors that influence the returns of the market. For example, in the Fama–French 3-factor model, $f_1$, $f_2$ and $f_3$ are respectively the excessive returns of the market portfolio, which is the value-weighted return on all NYSE, AMEX and NASDAQ stocks over the one-month Treasury bill rate, a portfolio constructed based on the market capitalization, and a portfolio constructed based on the book-to-market ratio. Of course, constructing factors that influence the market itself is a high-dimensional model selection problem with massive amount of trading data. The $K$-factor model [15], [74] assumes

$$Y_i = b_{i1} f_1 + \cdots + b_{iK} f_K + \varepsilon_i, \quad i = 1, \ldots, d, \tag{6.10}$$

where $\{\varepsilon_i\}$ are idiosyncratic noises, uncorrelated with the factors, and $d$ is the number of assets under consideration. This an extension of the famous Capital Asset Pricing Model derived by Sharpe and Lintner (See [10], [12]). Putting it into the matrix form, we have $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{\varepsilon}$ so that

$$\Sigma = \mathrm{Var}(\boldsymbol{B}\boldsymbol{f}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{B}\,\mathrm{Var}(\boldsymbol{f})\boldsymbol{B}^T + \Sigma_0, \tag{6.11}$$

where $\Sigma = \mathrm{Var}(\boldsymbol{y})$ and $\Sigma_0 = \mathrm{Var}(\boldsymbol{\varepsilon})$ is assumed to be diagonal.

Suppose that we have observed the returns of $d$ stocks over $n$ periods (e.g., 3 years daily data). Then, applying the least-squares estimate separately to each stock in (6.10), we obtain the estimates of coefficients in $\boldsymbol{B}$ and $\Sigma_0$. Now, estimating $\mathrm{Var}(\boldsymbol{f})$ by its sample variance, we obtain a substitution estimator $\widehat{\Sigma}$ using (6.11). On the other hand, we can also use the sample covariance matrix, denoted by $\widehat{\Sigma}_{\mathrm{sam}}$, as an estimator.

In the risk management or portfolio allocation, the number of stocks $d$ can be comparable with the sample size $n$ so it is better modeled as $d_n$. Fan *et al.* [32] investigated thoroughly when the estimate $\widehat{\Sigma}$ outperforms $\widehat{\Sigma}_{\mathrm{sam}}$ via both asymptotic and simulation studies. Let us quote some of their results.

**Theorem 7.** *Let $\lambda_k(\Sigma)$ be the $k$-th largest eigenvalue of $\Sigma$. Then, under some regularity conditions, we have*

$$\max_{1\le k\le d_n} |\lambda_k(\widehat{\Sigma}) - \lambda_k(\Sigma)| = o_P\{(\log n\, d_n^2/n)^{1/2}\} = \max_{1\le k\le d_n} |\lambda_k(\widehat{\Sigma}_{\mathrm{sam}}) - \lambda_k(\Sigma)|.$$

*For a selected portfolio weight $\boldsymbol{\xi}_n$ with $\mathbf{1}^T\boldsymbol{\xi}_n = 1$, we have*

$$\left|\boldsymbol{\xi}_n^T\widehat{\Sigma}\boldsymbol{\xi}_n - \boldsymbol{\xi}_n^T\Sigma\boldsymbol{\xi}_n\right| = o_P\{(\log n\, d_n^4/n)^{1/2}\} = \left|\boldsymbol{\xi}_n^T\widehat{\Sigma}_{\text{sam}}\boldsymbol{\xi}_n - \boldsymbol{\xi}_n^T\Sigma_n\boldsymbol{\xi}_n\right|.$$

*If, in addition, the all elements in $\boldsymbol{\xi}_n$ are positive, then the latter rate can be replaced by $o_P\{(\log n\, d_n^2/n)^{1/2}\}$.*

The above result shows that for risk management where the portfolio risk is $\boldsymbol{\xi}_n^T\Sigma\boldsymbol{\xi}_n$, no substantial gain can be realized by using the factor model. Indeed, there is no substantial gain for estimating the covariance matrix even if the factor model is correct. These have also convincingly been demonstrated in [32] using simulation studies. Fan *et al.* [32] also gives the order $d_n$ under which the covariance matrix can be consistently estimated.

The substantial gain can be realized if $\Sigma^{-1}$ is estimated. Hence, the factor model can be used to improve the construction of the optimal mean-variance portfolio, which involves the inverse of the covariance matrix. Let us quote one theorem of [32]. See other results therein for optimal portfolio allocation.

**Theorem 8.** *Under some regularity conditions, if $d_n = n^\alpha$, then for $0 \le \alpha < 2$,*

$$d_n^{-1}\text{tr}(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I_{d_n})^2 = O_P(n^{-2\beta})$$

*with $\beta = \min(1/2, 1-\alpha/2)$, whereas for $\alpha < 1$, $d_n^{-1}\text{tr}(\Sigma^{-1/2}\widehat{\Sigma}_{\text{sam}}\Sigma^{-1/2} - I_{d_n})^2 = O_P(d_n/n)$. In addition, under the Frobenius norm*

$$d_n^2\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|^2 = o(d_n^4\log n/n) = \|\widehat{\Sigma}_{\text{sam}}^{-1} - \Sigma^{-1}\|^2.$$

**6.4. Machine learning and data mining.** In machine learning, our goal is to build a model with the capability of good prediction of future observations. Prediction error depends on the loss function, which is also referred to as a divergence measure. Many loss functions are used in the literature. To address the versatility of loss functions, let us use the device introduced by [7]. For a concave function $q(\cdot)$, define a $q$-class of loss function $\ell(\cdot,\cdot)$ to be

$$\ell(y,\hat{m}) = q(\hat{m}) - q(y) - q'(\hat{m})(\hat{m} - y) \tag{6.12}$$

where $\hat{m} \equiv \hat{m}(\boldsymbol{x})$, an estimate of the regression function $m(\boldsymbol{x}) = E(y|\boldsymbol{x})$. Due to the concavity of $q$, $\ell(\cdot,\cdot)$ is non-negative.

Here are some notable examples of $\ell$-loss constructed from the $q$-function. For binary classification, $y \in \{-1, 1\}$. Letting $q(m) = 0.5\min\{1-m, 1+m\}$ yields the misclassification loss, $\ell_1(y,\hat{m}) = I\{y \ne I(\hat{m} > 0)\}$. Furthermore, $\ell_2(y,\hat{m}) = [1 - y\text{sgn}(\hat{m})]_+$ is the hinge loss if $q(m) = \frac{1}{4}\min\{1-m, 1+m\}$. The function $q_3(m) = \sqrt{1-m^2}$ results in $\ell_3(y,\hat{m}) = \exp\{-0.5y\log\{(1+\hat{m})/(1-\hat{m})\}$, the exponential loss function in AdaBoost [40]. Taking $q(m) = cm - m^2$ for some constant $c$ results in the quadratic loss $\ell_4(y,\hat{m}) = (y-\hat{m})^2$.

For a given loss function, we may extend the PLS to a penalized empirical risk minimization (4.5). The dimensionality $d$ of the feature vectors can be much larger than $n$ and hence the penalty is needed to select important feature vectors. See, for example, [4] for an important study in this direction.

We next make a connection between the penalized loss function and the popularly used support vector machines (SVMs), which have been successfully applied to various classification problems. In binary classification problems, the response $y$ takes values either 1 or $-1$, the class labels. A classification rule $\delta(\boldsymbol{x})$ is a mapping from the feature vector $\boldsymbol{x}$ to $\{1, -1\}$. Under the 0–1 loss, the misclassification error of $\delta$ is $P\{y \neq \delta(\boldsymbol{x})\}$. The smallest classification error is the Bayes error achieved by $\text{argmin}_{c\in\{1,-1\}} P(y = c|\boldsymbol{x})$. Let $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \ldots, n$ be a set of training data, where $\boldsymbol{x}_i$ is a vector with $d$ features, and the output $y_i \in \{1, -1\}$ denotes the class label. The 2-norm SVM is to find a hyperplane $\boldsymbol{x}^T\boldsymbol{\beta}$, in which $x_{i1} = 1$ is an intercept and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{(2)}^T)^T$, that creates the biggest margin between the training points from class 1 and $-1$ [85]:

$$\max_{\boldsymbol{\beta}} \frac{1}{\|\boldsymbol{\beta}_{(2)}\|^2} \quad \text{subject to } y_i(\boldsymbol{\beta}^T\boldsymbol{x}_i) \geq 1 - \xi_i, \text{ for all } i, \xi_i \geq 0, \sum \xi_i \leq B, \quad (6.13)$$

where $\xi_i$ are slack variables, and $B$ is a pre-specified positive number that controls the overlap between the two classes. Due to its elegant margin interpretation and highly competitive performance in practice, the 2-norm SVM has become popular and has been applied for a number of classification problems. It is known that the linear SVM has an equivalent hinge loss formulation [48]

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} [1 - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta})]_+ + \lambda \sum_{j=2}^{d} \beta_j^2.$$

Lin [62] shows that the SVM directly approximates the Bayes rule without estimating the conditional class probability because of the unique property of the hinge loss. As in the ridge regression, the $L_2$-penalty helps control the model complexity to prevent over-fitting.

Feature selection in the SVM has received increasing attention in the literature of machine learning. For example, the last issue of volume 3 (2002-2003) of *Journal of Machine Learning Research* is a special issue on feature selection and extraction for SVMs. We may consider a general penalized SVM

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} [1 - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta})]_+ + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

The 1-norm (or LASSO-like) SVM has been used to accomplish the goal of automatic feature selection in the SVM ([89]). Friedman *et al.* [41] shows that the 1-norm SVM is preferred if the underlying true model is sparse, while the 2-norm SVM performs better if most of the predictors contribute to the response. With the SCAD penalty, the penalized SVM may improve the bias properties of the 1-norm SVM.

## References


[1] Antoniadis, A., and Fan, J., Regularization of wavelets approximations (with discussions). *J. Amer. Statist. Assoc.* **96** (2001), 939–967.

[2] Benjamini, Y., and Yekutieli, D., The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** (2001), 1165–1188.

[3] Bickel, P. J., Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (ed. by M. H. Rizvi, J. S. Rustagi, and D. Siegmund), Academic Press, New York 1983, 511–528.

[4] Bickel, P. J., and Levina, E., Some theory for Fisher's linear discriminant, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** (2004), 989–1010.

[5] Brown, P. O., and Botstein, D., Exploring the new world of the genome with microarrays. *Nat. Genet.* **21** (suppl. 1) (1999), 33–37.

[6] de Boor, C., *A Practical Guide to Splines*. Appl. Math. Sci. 27, Springer-Verlag, New York 1978.

[7] Bregman, L. M., A relaxation method of finding a common points of convex sets and its application to the solulation of problems in convex programming. *U.S.S.R. Comput. Math. Math. Phys.* **7** (1967), 620–631.

[8] Breiman, L., Better subset regression using the nonnegative garrote. *Technometrics* **37** (1995), 373–384.

[9] Cai, J., Fan, J., Li, R., and Zhou, H., Variable selection for multivariate failure time data. *Biometrika* **92** (2005), 303–316.

[10] Campbell, J. Y., Lo, A., and MacKinlay, A. C., *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ, 1997.

[11] Craven, P., and Wahba, G., Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** (1979), 377–403.

[12] Cochrane, J. H., *Asset Pricing*. Princeton University Press, Princeton, NJ, 2001.

[13] Coifman, R. R., and Saito, N., Constructions of local orthonormal bases for classification and regression. *C. R. Acad. Sci. Paris Sér. I Math.* **319** (1994), 191–196.

[14] Coifman, R. R., and Wickerhauser, M. V., Entropy-based algorithms for best-basis selection. *IEEE Trans. Inform. Theory* **38** (1992), 713–718.

[15] Chamberlain, G., and Rothschild, M., Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** (1983), 1281–1304.

[16] Diaconis, P., and Sturmfels, B., Albegraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** (1998), 363–397.

[17] Daubechies, I., *Ten Lectures on Wavelets*. SIAM, Philadelphia 1992.

[18] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B* **39** (1977), 1–38

[19] Donoho, D. L., High-dimensional data analysis: the curses and blessings of dimensionality. *Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Centrury* (2000). Available at http://www-stat.stanford.edu/~donoho/Lectures.

[20] Donoho, D. L., and Elad, E., Maximal sparsity representation via $l_1$ Minimization. *Proc. Nat. Aca. Sci.* **100** (2003), 2197–2202.

[21] Donoho, D. L., and Huo, X., Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), 2845–2862.

[22] Donoho, D. L., and Jin, J., Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** (2004), 962–994.

[23] Donoho, D. L., and Johnstone, I. M., Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** (1994), 425–455.

[24] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D., Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc. B* **57** (1995), 301–369.

[25] Dudoit, S., Shaffer, J. P., and Boldrick, J. C., Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** (2003), 71–103.

[26] Dudoit, Y., Yang, Y. H., Callow, M. J., and Speed, T. P., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12** (2002), 111–139.

[27] Efron, B., Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** (2004), 96–104.

[28] Eilers, P. H. C., and Marx, B. D., Flexible smoothing with $B$-splines and penalties. *Statist. Sci.* **11** (1996), 89–121.

[29] Fan, J., A selective overview of nonparametric methods in financial econometrics (with discussion). *Statist. Sci.* **20** (2005), 316–354.

[30] Fan, J., Chen, Y., Chan, H. M., Tam, P., and Ren, Y., Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells. *Proc. Natl Acad. Sci. USA* **103** (2005), 17751–17756.

[31] Fan, J., and Gijbels, I. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London 1996.

[32] Fan, J., Fan, Y., and Lv, J., Large dimensional covariance matrix estimation using a factor model. Manuscript, 2005.

[33] Fan, J., and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** (2001), 1348–1360.

[34] Fan, J., and Li, R., Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** (2002), 74–99.

[35] Fan, J., and Peng, H., Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** (2004), 928–961.

[36] Fan, J., Peng, H., and Huang, T., Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency (with discussion). *J. Amer. Statist. Assoc.* **100** (2005), 781–813.

[37] Fan, J., Tam, P., Vande Woude, G., and Ren, Y., Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl Acad. Sci. USA* **101** (2004), 1135–1140.

[38] Foster, D. P., and George, E. I., The risk inflation criterion for multiple regression. *Ann. Statist.* **22** (1994), 1947–1975.

[39] Frank, I. E., and Friedman, J. H., A statistical view of some chemometrics regression tools. *Technometrics* **35** (1993), 109–148.

[40] Freund, Y., Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Systems Sci.* **55** (1997), 119–139.

[41] Friedman, J., Hastie, T. Rosset, S., Tibshirani, R. and Zhu, J., Discussion of boosting papers. *Ann. Statist.* **32** (2004), 102–107.

[42] Greenshtein, E., Best subset selection, persistence in high-dimensional statistical learning and optimization under $\ell_1$-constraint. *Ann. Statist.* **34** (5) (2006), to appear.

[43] Greenshtein, E., and Ritov, Y., Persistence in high-dimensional predictor selection and the virtue of overparametrization. *Bernoulli* **10** (2004), 971–988.

[44] Genovese, C., and Wasserman, L., A stochastic process approach to false discovery control. *Ann. Statist.* **32** (2004), 1035–1061.

[45] Hall, P., Edgeworth expansion for Student's $t$ statistic under minimal moment conditions. *Ann. Probab.* **15** (1987), 920–931.

[46] Hall, P., Some contemporary problems in statistical sciences. *The Madrid Intelligencer* (2006), to appear.

[47] Hand, D. J., Mannila, H., and Smyth, P., *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

[48] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Ser. Statist., Springer-Verlag, New York 2001.

[49] Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L., Covariance selection and estimation via penalised normal likelihood. *Biometrika* (2006), 85–98.

[50] Huang, J., Wang, D., and Zhang, C., A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. *J. Amer. Statist. Assoc.* **100** (2005), 814–829.

[51] Hull, J. *Options, Futures, and Other Derivatives*. 5th ed., Prentice Hall, Upper Saddle Review, NJ, 2003.

[52] Hunter, D. R., and Li, R., Variable selection using MM algorithms. *Ann. Statist.* **33** (2005), 1617–1642.

[53] Jing, B. Y., Shao, Q.-M., and Wang, Q. Y., Self-normalized Cramér type large deviations for independent random variables. *Ann. Probab.* **31** (2003), 2167–2215.

[54] Johnstone, I. M., On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** (2001), 295–327.

[55] Korosok, M. R., and Ma, S., Marginal asymptotics for the "large $p$, small $n$" paradigm: With applications to micorarray data. *Ann. Statist.*, to appear.

[56] Lange, K., A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statist. Soc. B* **57** (1995), 425–437.

[57] Lehmann, E. L., Romano, J. P., and Shaffer, J. P., On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33** (2005), 1084–1108.

[58] Li, H., and Gui, J., Gradient directed regularization for sparse Gaussian concentration graphs, with applications of inference of genetic networks. *Biostatistics* (2006), to appear.

[59] Li, R., Dziak, J., and Ma, H. Y., Nonconvex penalized least squares: characterizations, algorithm and application. Manuscript, 2006.

[60] Li, R., Root, T., and Shiffman, S., A local linear estimation procedure for functional multilevel modeling. In *Models for Intensive Longitudinal Data* (ed. by T. Walls and J. Schafer), Oxford University Press, New York 2006, 63–83.

[61] Lipschutz, R. J., Fodor, S., Gingeras, T., Lockhart, D. J., High density synthetic oloigonucleotide arrays. *Nat. Genet.* **21** (1999), 20–24.

[62] Lin, Y., Support vector machine and the Bayes rule in classification. *Data Mining and Knowledge Discovery* **6** (2002), 259–275.

[63] Ma, S, Kosorok, M. R., Huang, J., Xie, H., Manzella, L., and Soares, M. B., Robust semiparametric cDNA microarray normalization and significance analysis. *Biometrics* (2006), to appear.

[64] Mallat, S. G., Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* **315** (1989a), 69–87.

[65] Mallat, S. G., A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.* **11** (1989b), 674–693.

[66] Meinshausen, N., Lasso with relaxation. Manuscript, 2005.

[67] Meinshausen, N., and Bühlmann, P., High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** (3) (2006), to appear.

[68] Meyer, Y., *Ondelettes*. Hermann, Paris 1990.

[69] Miller, A. J., *Subset Selection in Regression*. Chapman&Hall/CRC, London 2002.

[70] Moffatt, H. K., *Risk Management: Value at Risk and Beyond*. Cambridge University Press, New York 2003.

[71] Nikolova, M., Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* **61** (2000), 633–658.

[72] Pachter, L., and Sturmfels, B., Parametric inference for biological sequence analysis. *Proc. Natl. Acad. Sci. USA* **101** (2004), 16138–16143.

[73] Pistone, G., Riccomagno, E., and Wynn, H. P., *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall / CRC, London 2000.

[74] Ross, S., The arbitrage theory of capital asset pricing. *J. Economic Theory* **13** (1976), 341–360.

[75] Ruppert, D., Wand, M. P., Carroll, R. J., *Semiparametric regression*. Cambridge University Press, Cambridge 2003.

[76] Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K., Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** (1997), 1371–1470.

[77] Storey J. D., Taylor J. E., and Siegmund D., Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Royal Statist. Soc. B* **66** (2004), 187–205.

[78] Sturmfels, B., and Sullivant, S., Toric ideals of phylogenetic invariants. *J. Comput. Biol.* **12** (2005), 204–228

[79] Svrakic, N. M., Nesic, O., Dasu, M. R. K., Herndon, D., and Perez-Polo, J. R., Statistical approach to DNA chip analysis. *Recent Prog. Hormone Res.* **58** (2003), 75–93.

[80] Tai, Y. C., and Speed, T. P., A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* **34** (5) (2006), to appear.

[81] Tibshirani, R., Regression shrinkage and selection via the LASSO. *J. Royal Statist. Soc. B* **58** (1996), 267–288.

[82] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.* **18** (2003), 104–117.

[83] Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H., Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29** (2001), 2549–2557.

[84] Tusher, V. G., Tibshirani, R., and Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** (2001), 5116–5121.

[85] Vapnik, V., *The Nature of Statistical Learning*. Springer-Verlag, New York 1995.

[86] Walls, T., and Schater, J., *Models for Intensive Longitudinal Data*, Oxford University Press, New York 2006.

[87] Wang, Y., Selective review on wavelets in Statistics. In *Frontiers of Statistics* (ed. by J. Fan and H. Koul), Imperial College Press, 2006.

[88] Zhang, H. P., Yu, C. Y., and Singer, B., Cell and tumor classification using gene expression aata: Construction of forests. *Proc. Natl. Acad. Sci. USA* **100** (2003), 4168–4172.

[89] Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R., 1-norm support vector machines. *Neural Information Processing Systems* **16** (2003).

[90] Zou, H., The adaptive Lasso and its oracle properties. Manuscript, 2005.

Department of Operation Research and Financial Engineering, and Bendheim Center for Finance, Princeton University, Princeton, NJ 08544, U.S.A.
E-mail: jqfan@Princeton.edu

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, U.S.A.
E-mail: rli@stat.psu.edu

# Random matrices and enumeration of maps

Alice Guionnet

**Abstract.** We review recent developments in random matrix theory related with the enumeration of connected oriented graphs called maps. In particular, we show that the long standing use of matrix integrals in physics to tackle such issues can be made rigorous and discuss some applications. This talk is based on joint works with E. Maurel-Segala and O. Zeitouni.

**Mathematics Subject Classification (2000).** Primary 15A52, 05C30.

**Keywords.** Random matrices, map enumeration.

## 1. Introduction

A map is a connected oriented diagram which can be embedded into a surface. Its genus $g$ is by definition the smallest genus of a surface in which it can be embedded in such a way that edges do not cross and the faces of the graph (which are defined by following the boundary of the graph) are homeomorphic to a disc. One has the formula for the Euler characteristic $\chi$:

$$\chi = 2 - 2g = \sharp \text{ vertices} + \sharp \text{ faces} - \sharp \text{ edges}.$$

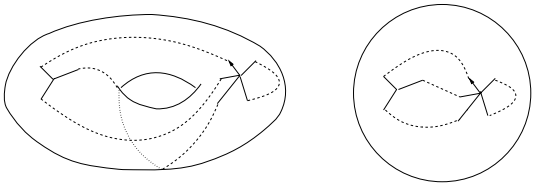


Figure 1. Examples of maps with 2 vertices of degree 3 and 5 respectively, with $g = 1$ and $g = 0$.

In the sequel, we shall be interested in the enumeration of maps up to equivalence classes, namely up to homeomorphisms of the oriented surface. This amounts to consider the purely combinatorial problem of enumerating the possible arrangements of the edges of graphs with prescribed vertices and genus. A dual point of view goes as follows. We can replace each vertex with valence $k$ by a face whose boundary is made of $k$ edges (each of them crossing a different edge adjacent to the vertex). The problem can then be reformulated as the enumeration of possible tilings of a surface

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of given genus by a given number of faces with prescribed degree (the degree of the face is the number of edges that border the face). For some problems (for instance related with statistical mechanics), one would like eventually to color the edges or the vertices of the map and impose additional constraints for the gluing of edges/vertices of different colors.

The problem of enumerating maps was first tackled in the sixties by W. Tutte [34], [35] who was motivated by combinatorial problems such as the four color problem (see [26] or [4] for combinatorial motivations and problems). Tutte considered *rooted* planar maps. The root of a map is a distinguished oriented edge. Fixing a root allows to reduce the number of symmetries of the problem; enumerating rooted maps is equivalent to count maps with labelled edges. Tutte showed that diverse 'chirurgical' operations on rooted planar maps allow to obtain equations for the generating functions of the numbers of these maps with a given number of faces, each face having the same fixed degree. One of the examples of the maps which were exactly enumerated by Tutte are triangulations (i.e maps with faces of degree 3); he proved [34] that the number of rooted triangulations with $2n$ faces is given by $2^{n+1}(3n)!/(2n+2)!n!$ (see e.g. E. Bender and E. Canfield [5] for generalizations). In general, the equations obtained by Tutte's approach are not exactly solvable; their analysis was the subject of subsequent developments (see e.g. [19]).

Because this last problem is in general difficult, a bijective approach was developed after the work of R. Cori and B. Vauquelin [13] and G. Schaeffer's thesis (see e.g. [32]). It was shown that planar triangulations and quadrangulations can be encoded by labelled trees, which are much easier to count. This idea proved to be very fruitful in many respects. It allows not only to study the number of maps but also part of their geometry; P. Chassaing and G. Schaeffer [12] could prove that the diameter of uniformly distributed quadrangulations with $n$ vertices behaves like $n^{\frac{1}{4}}$. This technique was first applied to triangulations or quadrangulations, but soon generalized to other maps, see e.g. [10] or [9]. The case of planar bi-colored maps related to the so-called Ising model on random planar graphs could also be studied [7]. Further, it allows also to tackle maps with higher genus, an avenue recently opened by M. Marcus and G. Schaeffer. In general, this approach give more complete results than the other methods. However, it yet can not cover all the models which were analysed in physics by the so-called matrix models approach and when it does, the solution for the enumeration problem has the same flavour than the solution obtained with matrix models (see [9]).

The question of enumerating maps has been studied intensively in physics for more than thirty years. One of the first motivation came in QCD (which stands for Quantum Chromodynamics) with a large number $N$ of colors; 't Hooft [33] noticed in the seventies that as $N$ is large, physical quantities can be expanded, via Feynamnn diagrams, as sums over maps. This fundamental remark allowed the connection between quantum field theory and the problem of enumerating maps, and in particular led to the use of matrix integrals to count maps (In [11], this technique was used to

enumerate planar maps with vertices of degree 4). The interest in enumerating maps was revived by quantum gravity in the eighties; random triangulations could be used for instance to approximate fluctuating geometries. As a side product, people got interested by statistical models defined on random graphs. Such models should in fact be related at criticality with the corresponding model on $\mathbb{Z}^2$ (see [28]). Maps were also used to approximate low-dimensional string theory (see e.g. the review [15]). Although recently the methods introduced by R. Cori, B. Vauquelin and G. Scheaffer began to be developed in physics too (by P. Di Francesco et al.), the most common approach has been to use matrix models, a rather indirect but quite powerful method that we shall describe in this survey (see also A. Zvonkin [36]). It is based on the particular form of Gaussian moments as given by Wick formula; if $(G_1, \ldots, G_{2n})$ is a centered Gaussian vector, then Wick formula asserts that

$$\mathbb{E}[G_1 G_2 \ldots G_{2n}] = \sum_{\substack{1 \le s_1 < s_2 \ldots < s_n \le 2n \\ r_i > s_i}} \prod_{j=1}^{n} \mathbb{E}[G_{s_j} G_{r_j}].$$

Alternatively, this formula can be represented by Feynman diagrams. Let us now consider matrices from the Gaussian Unitary Ensemble (GUE). For a fixed dimension $N$, let $\mathcal{H}_N$ be the set of $N \times N$ Hermitian matrices. The law of the GUE is then given as the Gaussian law on $\mathcal{H}_N$

$$\mu_N(dA) = \frac{1}{Z_N} e^{-\frac{N}{2}\operatorname{tr}(A^2)} dA.$$

In other words, $A_{lk} = \bar{A}_{kl}$ for $1 \le k < l \le N$ and

$$A_{kl} = (2N)^{-\frac{1}{2}}(g^1_{kl} + i g^2_{kl}) \quad \text{for } k < l, \qquad A_{kk} = N^{-\frac{1}{2}} g^1_{kk},$$

where the $(g^1_{kl}, g^2_{kl}, k \le l)$ are independent identically distributed standard Gaussian variables. One can then observe that Wick formula implies that for all integer numbers $p_i$, $1 \le i \le k$, all $k \in \mathbb{N}$, $N \in \mathbb{N}$,

$$\int \prod_{i=1}^{k} (N \operatorname{tr}(A^{p_i}))\, d\mu_N(A) = \sum_{F \ge 0} N^{F+k-\frac{\sum p_i}{2}}\, G((p_i)_{1 \le i \le k}, F) \tag{1}$$

with

$$G((p_i)_{1 \le i \le k}, F) = \sharp\{\text{oriented graphs with } F \text{ faces and } 1 \text{ vertex of degree } p_i, 1 \le i \le k\}.$$

In $G((p_i)_{1 \le i \le k}, F)$, the edges of the graph are labelled. One should notice that the number $F + k - \frac{\sum p_i}{2}$ corresponds to $2 - 2g$, with $g$ the genus of the surface on which a connected oriented graph with $F$ faces and one vertex of degree $p_i$ for $1 \le i \le k$

can be embedded, since such a graph has $k$ vertices and $2^{-1}\sum p_i$ edges. Hence, if we see the dimension $N$ of the matrices as a parameter, the expectation of the trace of moments of matrices from the GUE can be seen as a generating function for the number of oriented graphs with a given genus and a given number of vertices with prescribed degree. Laplace transforms of traces of matrices from the GUE should therefore be generating functions for maps. In fact, we find, by expanding the exponential and using (1) that, with $\boldsymbol{t} = (t_1, \dots, t_k)$,

$$\log Z_N(\boldsymbol{t}) := \log \int e^{-\sum_{i=1}^k t_i N \operatorname{tr}(A^{p_i})}\, d\mu_N(A) = \sum_{g\geq 0} N^{2-2g} F_g(\boldsymbol{t}) \tag{2}$$

with

$$F_g(\boldsymbol{t}) := \sum_{n_1,\dots,n_k\in\mathbb{N}^k} \prod_{i=1}^k \frac{(-t_i)^{k_i}}{k_i!} M((p_i, n_i)_{1\leq i\leq k}; g)$$

the generating function for the number $M((p_i, n_i)_{1\leq i\leq k}; g)$ of maps with genus $g$ and $n_i$ vertices of degree $p_i$ for $i \in \{1, \dots, k\}$. Note here that we now count maps, and so oriented graphs that are connected, due to the fact that we took the logarithm. Formula (2) is only formal, i.e means that all the derivatives of the functions on each side of the equality match at $(t_1, \dots, t_k) = (0, \dots, 0)$.

An interesting feature of the relation (1) is that it can be generalized to several matrices, corresponding then to the enumeration of colored-edges maps. Namely, let us introduce a bijection between non-commutative monomials and oriented vertices with colored half-edges and a distinguished half-edge as follows; to the letters $(X_1, \dots, X_m)$ we associate half-edges with $m$ different colors $c_1, \dots, c_m$, and to a monomial $q(X_1, \dots, X_m) = X_{i_1} \dots X_{i_k}$ a clockwise oriented vertex with first half-edge (which is distinguished) of color $c_{i_1}$, second of color $c_{i_2}$ till the last half-edge of color $c_{i_k}$. We call such a vertex, equipped with its colored half-edges, orientation and distinguished edge, a star of type $q$. It defines a bijection between monomials and stars. We then can generalize (2) as follows; let $(q_1, \dots, q_k)$ be $k$ non-commutative monomials of $m$ indeterminates, then

$$\begin{aligned}\int \prod_{i=1}^k (N \operatorname{tr}(q_i(A_1, \dots, A_m))\, d\mu_N(A_1) \dots d\mu_N(A_m) \\ = \sum_{F\geq 0} N^{k+F-\frac{\sum p_i}{2}} G_c((q_i)_{1\leq i\leq k}, F),\end{aligned} \tag{3}$$

with $G_c((q_i)_{1\leq i\leq k}, F)$ the number of oriented graphs with $F$ faces and one star of type $q_i$, $1 \leq i \leq k$, the gluing between half-edges of different colors being forbidden. (2) also generalizes to this multi-matrix setting and we find that

$$\log Z_N(\boldsymbol{t}) = \log \int e^{-N\sum_{i=1}^k t_i \operatorname{tr}(q_i(A_1,\dots,A_m))} \prod_{i=1}^m d\mu_N(A_i) \tag{4}$$

expends formally as a generating function of colored maps (here $M((p_i, n_i), 1 \leq i \leq k; g)$ has to be replaced by the number $M_c((q_i, n_i), 1 \leq i \leq k; g)$ of maps with genus $g$ and $n_i$ stars of type $q_i$, the gluing between half-edges of different colors being forbidden.)

As we said before, these considerations have been intensively used in physics to analyze various combinatorial models via their representations in terms of matrices. It is no surprise that mathematicians end up wondering what physicists are doing or come to cross the same lines of thoughts. In the last ten years, progress in the theory of random matrices led to a better mathematical understanding of this approach. The first natural question is to find a reasonable domain of the parameters $(t_1, \ldots, t_n)$ where the expansion (2) or (4) are not only formal. Because the right hand side is a priori a diverging series, this expansion can not be obtained analytically in a neighborhood of the origin, but we would like to show that equality holds up to some error term $N^{-2k}$ provided the parameters belong to some neighborhood of the origin. Once this question is settled, one can try to 'solve'(and then in which sense?) the combinatorial problem by estimating the matrix integral.

As we shall see in the next section, the first goal has received a rather complete answer in the last few years. For the second question and one matrix setting, it turns out that at least the first order asymptotics of the left-hand-side of (2) can be computed by using standard saddle point (or large deviations) techniques. The answer is yet not very transparent since it is given by a variational formula and we shall review part of its analysis. In the multi-matrix setting, very few results have been obtained so far, a few of which we shall describe.

## 2. Expansion of the free energy of matrix models

**2.1. One matrix integrals.** In the case of one matrix, the free energy of the matrix model can be expressed as an integral over the eigenvalues of the random matrix. It is well known (see [30]) that the law of the eigenvalues of the GUE can be described by a Coulomb gas law;

$$d\sigma_N(\lambda_1, \ldots, \lambda_N) = \frac{1}{Z_N} \prod_{i<j} |\lambda_i - \lambda_j|^2 e^{-\frac{N}{2}\sum_{i=1}^N (\lambda_i)^2} \prod d\lambda_i$$

with $Z_N$ the normalizing constant. It therefore turns out that (2) is given by

$$Z_N(\boldsymbol{t}) = Z_N^{-1} \int e^{-N\sum_{i=1}^N [V_{\boldsymbol{t}}(\lambda_i) + \frac{1}{2}(\lambda_i)^2]} \prod_{i<j} |\lambda_i - \lambda_j|^2 \prod d\lambda_i \tag{5}$$

with $V_{\boldsymbol{t}}(x) = \sum_{i=1}^k t_i x^{p_i}$. To make sure that $Z_N(\boldsymbol{t})$ is finite for each $N \in \mathbb{N}$, we shall assume that $p_k = \max_{l \leq k} p_l$ is even and $t_k > 0$. It was conjectured by Bessis, Itzykson and Zuber [6] that $\log Z_N(\boldsymbol{t})$ can be expanded in the vicinity of the origin. It

was only twenty years later that this question met its complete mathematical treatment in [17] (see also [2], [1] for previous advances in the subject). It was indeed shown in [17], Theorem 1.1, that if we let

$$\mathbb{T}(T,\gamma)=\left\{\boldsymbol{t}\in\mathbb{R}^k : \sum_{i=1}^k |t_i|\le T,\ t_k>\gamma\sum_{i=1}^{k-1}|t_i|\right\},$$

then the following result holds:

**Theorem 2.1.** *For all $k\in\mathbb{N}$, there is $T>0$ and $\gamma>0$ so that for $\boldsymbol{t}\in\mathbb{T}(T,\gamma)$, for all $k\in\mathbb{N}$, one has the expansion*

$$N^{-2}\log Z_N(\boldsymbol{t})=\sum_{l=0}^{k}N^{-2l}F_l(\boldsymbol{t})+O(N^{-2k-2}).$$

*Moreover,*

$$F_l(\boldsymbol{t})=\sum_{n_1,\dots,n_k\in\mathbb{N}^n}\prod_{l=1}^{k}\frac{(-t_i)^{n_i}}{n_i!}M((p_i,n_i),1\le i\le k;g)$$

*with $M((p_i,n_i)_{1\le i\le k};g)$ the number of maps with genus $g$ and $n_i$ vertices of degree $p_i$ for $1\le i\le k$.*

This result is based on an expansion for the mean empirical density of the eigenvalues under the associated Gibbs measure

$$d\sigma_{\boldsymbol{t}}^N(\lambda_1,\dots,\lambda_N)=Z_N(\boldsymbol{t})^{-1}e^{-N\sum_{i=1}^N V_t(\lambda_i)}\prod_{i<j}|\lambda_i-\lambda_j|^2e^{-\frac{N}{2}\sum_{i=1}^N(\lambda_i)^2}\prod d\lambda_i.$$

Indeed, if we set $\mu_{\boldsymbol{t}}^N$ to be the probability measure on $\mathbb{R}$ given for any bounded measurable test function $f$ by

$$\int f(x)d\mu_{\boldsymbol{t}}^N(x)=\int\frac{1}{N}\sum_{i=1}^N f(\lambda_i)\,d\sigma_{\boldsymbol{t}}^N(\lambda_1,\dots,\lambda_N),$$

it is proved in [17], Theorem 1.3, that

**Theorem 2.2.** *For all $k\in\mathbb{N}$, there is $T>0$ and $\gamma>0$ so that for $\boldsymbol{t}\in\mathbb{T}(T,\gamma)$, for all $k\in\mathbb{N}$, one has the expansion*

$$\int f(x)d\mu_{\boldsymbol{t}}^N(x)=\sum_{g=0}^{k}N^{-2g}f_g(\boldsymbol{t})+O(N^{-2k-2})$$

*for any smooth function $f$ which grows no faster than a polynomial at infinity. Moreover, for all $p\in\mathbb{N}$, if $f(x)=x^p$,*

$$f_g(\boldsymbol{t})=\sum_{n_1,\dots,n_k\in\mathbb{N}^n}\prod_{l=1}^{k}\frac{(-t_i)^{n_i}}{n_i!}M((p,1),(p_i,n_i)_{1\le i\le k};g).$$

Note that this second theorem implies the first since for all $l$,

$$\partial_{t_l} \log Z_N(\boldsymbol{t}) = \mu_{\boldsymbol{t}}^N(x^{p_l})$$

gives

$$\log \frac{Z_N(\boldsymbol{t})}{Z_N(\boldsymbol{0})} = \sum_{l=1}^{k} \int_0^{t_l} \mu^N_{(0,\dots,0,s,t_{l+1},\dots,t_k)}(x^{p_l})\, ds.$$

The proof of these results are based on orthogonal polynomials; because the interaction between the eigenvalues $(\lambda_1, \dots, \lambda_N)$ are given in terms of the square of a Vandermonde determinant, the density of the law $\mu_{\boldsymbol{t}}^N$ can be expressed in terms of orthogonal polynomials, whose limits are well known (since they are completely integrable). The theory of integrable systems allowed many important breakthroughs in the theory of matrix models, but we want to argue in the next section that the large $N$ expansion of matrix models can be obtained by more direct arguments.

**2.2. Many matrix integrals.** In [22], [23], [29], following (3), we considered the multi-matrix integral defined, for $k$ non-commutative monomials of $m$ indeterminates $(q_1, \dots, q_k)$, by

$$Z_N(\boldsymbol{t}) = \int e^{-N \sum_{i=1}^k t_i \operatorname{tr}(q_i(A_1,\dots,A_m))}\, d\mu_N(A_1) \dots d\mu_N(A_m). \tag{6}$$

To make this integral finite and not oscillatory, we assume the following. Let $*$ be the involution on polynomial functions of $m$ non-commutative integrals given by

$$(zX_{i_1} \dots X_{i_p})^* = \bar{z} X_{i_p} \dots X_{i_1}$$

for any $p \in \mathbb{N}$ and any $i_j \in \{1, \dots, m\}$. Then, to avoid possible oscillations, we assume that $V_{\boldsymbol{t}}(X_1, \dots, X_m) = \sum_{i=1}^k t_i q_i(X_1, \dots, X_m)$ is self-adjoint, i.e. $V_{\boldsymbol{t}} = V_{\boldsymbol{t}}^*$. To bound the integral, we assume that there exists $c > 0$ so that $V_{\boldsymbol{t}}$ is $c$-convex, i.e $W(X_1, \dots, X_m) = V_{\boldsymbol{t}}(X_1, \dots, X_m) + \frac{(1-c)}{2} \sum_{i=1}^k X_i^2$ is convex in the sense that for any $N \in \mathbb{N}$, the application

$$(X_1, \dots, X_m) \in \mathcal{H}_N^m \to \operatorname{tr}(W(X_1, \dots, X_m))$$

is a convex function of the entries of $(X_1, \dots, X_m)$. Observe that, by Klein's lemma, if $V$ is a convex function of one real variable, $V$ is convex in the above sense and therefore our condition includes all $V_{\boldsymbol{t}}$ of the form

$$V_{\boldsymbol{t}}(X_1, \dots, X_m) = \sum V_i(\sum \kappa_i^j X_j) + \sum \beta_{jl} X_i X_j$$

with $V_i$ convex functions of one variable, real numbers $\kappa_i^j$ and $\beta_{jl}$ small enough constant (depending on $c$). This assumption generalizes that of Theorem 2.2 since if $\gamma$ is large enough and $T$ small enough, for $\boldsymbol{t} \in \mathbb{T}(T, \gamma)$, the potential $V_{\boldsymbol{t}}(x) =$

$\sum_{i=1}^k t_i x^{p_i} + 2^{-1}(1-c)x^2$ is strictly convex. In [22], [23], [29], the analogue of Theorems 2.1 and 2.2 were obtained for a range of parameters which are small enough and so that $V_{\boldsymbol{t}}$ stays uniformly $c$-convex for some $c > 0$. For the analogue of Theorem 2.2, $\mu_{\boldsymbol{t}}^N$ is generalized into the linear form on non-commutative polynomials given by

$$\mu_{\boldsymbol{t}}^N(P) = \frac{1}{Z_N(\boldsymbol{t})} \int \frac{1}{N} \operatorname{tr}(P(A_1, \ldots, A_m)) e^{-N \sum_{l=1}^k t_i \operatorname{tr}(q_i(A_1,\ldots,A_m))} \prod d\mu_N(A_i).$$

The techniques are complectly different from those of [17] (in fact, orthogonal polynomial techniques are unknown for general multi-matrix models) and rely on combinatorial interpretations of non-commutative differential operators. For instance, for the first order expansion, it can be shown under our hypothesis that for any non-commutative polynomial $P$, $\mu_{\boldsymbol{t}}^N(P)$ converges towards some quantity $\tau(P)$. Furthermore, $\tau$ satisfies some 'non-commutative differential equation', called the Schwinger–Dyson equation, which says that for all polynomials $P$,

$$\tau((X_i + D_i V_{\boldsymbol{t}})P) = \tau \otimes \tau(\partial_i P), \quad \tau(1) = 1 \tag{7}$$

with $\partial_i$ (resp. $D_i$) the non-commutative derivative (resp. the cyclic derivative) given on a monomial $P$ by

$$\partial_i P = \sum_{P = P_1 X_i P_2} P_1 \otimes P_2, \quad D_i P = \sum_{P = P_1 X_i P_2} P_2 P_1$$

where the sums run over all the possible decomposition of the monomial $P$ into $P_1 X_i P_2$. It turns out that in our range of parameters, there is only one solution to (7) (which satisfies some boundedness properties that the limit points of $\mu_{\boldsymbol{t}}^N$ share), which corresponds to the generating function for planar maps. This last identification comes out because $\partial_i$ and $D_i$ have very simple combinatorial interpretations; if you think of $\tau$ as the generating function of maps, you will see that $\partial_i$ consists in the operation of splitting your map into two disjoint maps when two edges of the color $c_i$ of one vertex are glued together, whereas $D_i$ will consist in the operation of erasing one edge when two different vertices are connected via an edge of the color $c_i$, then obtaining a single bigger vertex (see [22]). Amazingly, it turns out that these non-commutative derivatives play exactly the same role than the surgery initially introduced by Tutte. One could then wonder what matrix models brought so far. At least a funny remark; the limit $\tau$, whose moments are generating functions for maps, is a tracial state. In particular, in the one matrix case, it is a probability measure. Another remark is that the higher orders in the expansion of matrix integrals can be expressed in terms of $\tau$ and the differential operators defined by the $\partial_i$ and the $D_i$. Thus the expansion describes, without further thinking, the operations that one can do on a map of genus $g$ to enumerate it in terms of lower genus maps. In the next section, we shall therefore concentrate on planar maps and the analysis of the limiting state $\tau$.

## 3. Estimating matrix integrals

We shall focus in this section on the first order of matrix integrals, that is on planar maps.

**3.1. One matrix integrals.** It is easily seen by a saddle point method (or large deviations, see e.g. [3]) that with $Z_N(\boldsymbol{t})$ given by (5) and $V_{\boldsymbol{t}}(x) = \sum_{i=1}^k t_i x^{p_i}$, we have

$$\lim_{N\to\infty} \frac{1}{N^2} \log Z_N(\boldsymbol{t}) = \sup_{\mu\in\mathcal{P}(\mathbb{R})} \left\{ \int \log|x-y|\, d\mu(x) d\mu(y) - \int \left( V_{\boldsymbol{t}}(x) + \frac{x^2}{2} \right) d\mu(x) \right\} \tag{8}$$

up to a universal constant coming from the limit of $N^{-2}\log Z_N$. Moreover, the above supremum is achieved at a unique probability measure $\mu_{\boldsymbol{t}}$ and we have for all bounded continuous function,

$$\lim_{N\to\infty} \mu_{\boldsymbol{t}}^N(f) = \mu_{\boldsymbol{t}}(f).$$

In particular, at least for $\boldsymbol{t} \in \mathbb{T}(T,\gamma)$ as in Theorem 2.2, for all integer $p$, $\mu_{\boldsymbol{t}}(x^p)$ is a generating function for maps with one vertex of degree $p$ and so at least formally,

$$\begin{aligned} G\mu_{\boldsymbol{t}}(z) &:= \int \frac{1}{z-x}\, d\mu_{\boldsymbol{t}}(x) \\ &= \frac{1}{z} \sum_{l\geq 0} z^{-l} \sum_{n_1,\dots,n_k} \prod_{i=1}^k \frac{(-t_i)^{n_i}}{n_i!} M((l,1),(p_i,n_i)_{1\leq i\leq k};0) \end{aligned}$$

is a generating function for maps too. In [14] (see also [17]), the solution to the variational problem (8) has been studied. It turns out that in the small range of parameters we are considering, we have the following characterization of $\mu_{\boldsymbol{t}}$;

$$\mu_{\boldsymbol{t}}(dx) = \frac{1_{[\alpha(\boldsymbol{t}),\beta(\boldsymbol{t})]}}{2\pi} \sqrt{(x-\alpha(\boldsymbol{t}))(\beta(\boldsymbol{t})-x)}\, h_{\boldsymbol{t}}(x) dx$$

with $h_{\boldsymbol{t}}$ a polynomial given explicitly in terms of $V_{\boldsymbol{t}}$ and $\alpha(\boldsymbol{t})$, $\beta(\boldsymbol{t})$ determined by the set of equations

$$\int_{\alpha(\boldsymbol{t})}^{\beta(\boldsymbol{t})} \frac{(V_{\boldsymbol{t}}'(s)+s)}{\sqrt{(s-\alpha(\boldsymbol{t}))(\beta(\boldsymbol{t})-s)}} ds = 0, \quad \int_{\alpha(\boldsymbol{t})}^{\beta(\boldsymbol{t})} \frac{s(V_{\boldsymbol{t}}'(s)+s)}{\sqrt{(s-\alpha(\boldsymbol{t}))(\beta(\boldsymbol{t})-s)}} ds = 2\pi.$$

$\alpha(\boldsymbol{t})$, $\beta(\boldsymbol{t})$ are analytic functions of $\boldsymbol{t} \in \mathbb{T}(T,\gamma)$. This however does not give a very explicit formula for $G\mu_{\boldsymbol{t}}$. When $V_{\boldsymbol{t}}$ is a monomial, more detailed analysis were performed in [14]. It turns out that when $V_{\boldsymbol{t}}$ is even (see [9]), the analysis is more simple and, in the case $V_{\boldsymbol{t}} = tx^4$, can be pushed to obtain explicit formulas (see [11]).

**3.2. Many matrix integrals.** The problem of enumerating colored, or decorated, maps is much more challenging. In combinatorics, only the so-called Ising model on random quadrangulations could be tackled so far (see [7]). The list of models which could be 'solved' in physics is slightly longer; it includes for instance the so called Potts model, induced QCD, $ABAB$ model, dually weighted graphs (see e.g. [20] and references therein). Basically, all these models can be written in terms of quadratic interaction models, either by definition or by using character expansions. Thus, they are closely related with the Ising model we shall describe below. The Ising model is given by the partition function

$$Z_N(\boldsymbol{t}, c) = \int e^{-N\,\mathrm{tr}(V_{\boldsymbol{t}}^1(A))-N\,\mathrm{tr}(V_{\boldsymbol{t}}^2(B))-Nc\,\mathrm{tr}(AB)}\, d\mu_N(A) d\mu_N(B)$$

with $V_{\boldsymbol{t}}^1$ and $V_{\boldsymbol{t}}^2$ two polynomials of one real variable with coefficients depending on parameters $\boldsymbol{t}$. By paragraph 2.2, if $V_{\boldsymbol{t}}^1$, $V_{\boldsymbol{t}}^2$ are convex, for small enough parameters $(\boldsymbol{t}, c)$, the free energy $\log Z_N(\boldsymbol{t})$ expands into a generating function for two-colored maps with vertices prescribed by $V_{\boldsymbol{t}}^1$ and $V_{\boldsymbol{t}}^2$. The interaction $AB$ serves to generate edges between vertices of different colors. Thus, when $V_{\boldsymbol{t}}^1(x) = V_{\boldsymbol{t}}^2(x) = tx^4$, the model really looks like a generalization of the standard Ising model, with spins lying on a random quadrangulation rather than on $\mathbb{Z}^2$. Indeed, we have for small enough parameters $t, c$,

$$\frac{1}{N^2}\log Z_N(t, c) = (1-c^2)^{-1} \sum_{g=0}^{k} \frac{1}{N^{2g}} \sum_{k,\ell} \frac{1}{k!} \left(\frac{-t}{(1-c^2)^2}\right)^k \frac{(-c)^\ell}{\ell!} C(k, \ell, g) + o(N^{-2k})$$

with $C(k, \ell, g)$ the number of maps with genus $g$ with $k$ vertices of valence 4 being assigned the sign $+1$ or $-1$, with exactly $\ell$ edges between vertices of different signs. This formula is reminiscent of the standard grand canonical partition function for the Ising model in $\mathbb{Z}^2$, where $C(k, \ell, g)$ is simply replaced by the number of configurations on a subset of $\mathbb{Z}^2$ rather than on random graphs. The genus is then related with the boundary conditions. In this case, where $V_t^1(x) = V_t^2(x) = tx^4$, an explicit formula for the limiting free energy was obtained in [31] from which important information such as phase transition could be derived [8] (these results were recovered in [7] by a purely combinatorial approach). For more general potentials, variational formulas generalizing those of the one matrix setting were obtained in [20]. A more detailed analysis of these limits is under study. The basic ingredient for these general potentials estimates is based on the remark that under $\mu_N$, $A = UDU^*$ with $U$ a unitary matrix following the Haar measure and $D$ a diagonal matrix, independent of $U$. Therefore, the interaction in the Ising model is given by the spherical, or Itzykson–Zuber–Harish-Chandra, integral

$$I(D_1, D_2) = \int e^{Nc\,\mathrm{tr}(D_1 U D_2 U^*)}\, dU$$

with $D_1$ and $D_2$ the diagonal matrices of the eigenvalues of $A$, $B$ and $dU$ the Haar measure on the set of $N \times N$ unitary matrices. In [25], we obtained the large $N$ asymptotics of spherical integrals. Namely, take a sequence of diagonal matrices $(D_1^N, D_2^N)$ so that the empirical measures $N^{-1}\sum_{i=1}^N \delta_{D_j^N(ii)}$ converges weakly towards a probability measures $\mu_j$ for $j = 1, 2$. Then, if we set

$$I(\mu) = -\frac{1}{2}\iint \log|x-y| d\mu(x) d\mu(y) + \frac{1}{2}\int x^2 d\mu(x),$$

we have, if $I(\mu_1) < \infty$, $I(\mu_2) < \infty$,

$$\lim_{N\to\infty} \frac{1}{N^2} \log I(D_1^N, D_2^N)$$
$$= -\frac{1}{2} \inf \left\{ \int_0^1 \int u_t(x)^2 \rho_t(x) dx dt + \frac{\pi^2}{3} \int_0^1 \int \rho_t(x)^3 dx dt \right\} + I(\mu_1) + I(\mu_2).$$

The above infimum is taken over $(\rho, u)$ on $(0, 1) \times \mathbb{R}$ so that $v_t(dx) = \rho_t(x)dx$ is a probability measure on $\mathbb{R}$ for all $t \in (0, 1)$, $t \to v_t$ is continuous with limit as t goes to zero (resp. one) given by $\mu_1$ (resp. $\mu_2$) and for all $t \in (0, 1)$, all $x \in \mathbb{R}$,

$$\partial_t \rho_t(x) + \partial_x(\rho_t(x) u_t(x)) = 0.$$

In [20] it was shown that the infimum is taken at a couple $(\rho, u)$ so that $f = u + i\pi\rho$ satisfies the complex Burgers equation

$$\partial_t f_t(x) + f_t(x)\partial_x f_t(x) = 0.$$

These formulae are proved by large deviation estimates for $N$ non-intersecting Brownian motions evaluated at extremely small time $N^{-1}$. Complex Burgers equation also appears in discrete analogous settings coming from tiling, see e.g. [27]. Spherical integrals are rather fundamental objects since they are related with the characters of the symmetric group; the above limits give asymptotics of Schur functions, cf. [21].

## 4. Conclusion: Matrix models input in combinatorics

Even though rather indirect, the matrix model approach to the enumeration of maps have proved to be powerful since it permits to consider quite general maps. For general type of vertices, the formulas obtained by this method are often not so much explicit, but this should be no surprise. When possible, the bijective approach provides more detailed information, such as the diameter of the graph with a given number of vertices (an information which was never grasped by the matrix model approach so far). The matrix model approach shows that some (maybe) unexpected tools can be used to solve these combinatorial problems; let us cite characters expansions (see [24] and references therein), Brownian motion and stochastic calculus (see [25]). We believe

also that the description of the generating functions $\mu_t(x^p)$ as the expectation under a probability measure, or tracial state, is a rather powerful remark. It allows to give some information on Tutte's solutions to the equations for generating functions of maps, seen as a solution to Schwinger–Dyson's equation (7). For instance, it was shown in [16] that the generating function $\tau(z - A)^{-1}$ for the Ising model with general polynomial potentials satisfies an algebraic equation.

This field has experienced quite a lot of developments in the last few years, attracting the interests of theoretical physicists and of mathematicians from diverse fields such as combinatorics, integrable systems or probability. Central in the problem of the enumeration of maps is the Schwinger–Dyson equation (7) which encodes most of the induction relations satisfied by the numbers of interest. The study of its solution is the heart of the problem of enumerating maps and, at least in the multi-matrix model, also attracted the attention of free probabilists. Indeed, in free probability, the reverse question (i.e. given a tracial state, find a potential $V_t$ so that (7) is satisfied) serves to define the so-called conjugate variables which are central in free entropy questions. However, most issues in free probability are not related with small perturbative potentials as considered in this survey but on the contrary with very strong potentials. The understanding of matrix models is then extremly limited, since even the question of the convergence of the free energy is still unsettled.

## References

[1] Albeverio S., Pastur L., Shcherbina, M., On the $1/n$ expansion for some unitary invariant ensembles of random matrices. *Comm. Math. Phys.* **224** (2001), 271–305.

[2] Ambjørn, J., Chekhov, L., Kristjansen, C. F., and Makeenko, Yu., Matrix model calculations beyond the spherical limit. *Nuclear Phys. B* **404** (1993), 127–172; Erratum ibid. **449** (1995), 681.

[3] Ben Arous, G., Guionnet, A., Large deviations for Wigner's law and Voiculescu's non-commutative entropy. *Probab. Theory Related Fields* **108** (1997), 517–542.

[4] Bender, E., Some unsolved problems in map enumeration. *Bull. Inst. Combin. Appl.* **3** (1991), 51–56.

[5] Bender, E., Canfield, E., The number of degree-restricted rooted maps on the sphere. *SIAM J. Discrete Math.* **7** (1994), 9–15.

[6] Bessis, D., Itzykson, C., Zuber, J. B., Quantum field theory techniques in graphical enumeration. *Adv. in Appl. Math.* **1** (1980), 109–157.

[7] Bousquet-Melou, M., Schaeffer, G., The degree distribution in bipartite planar maps: applications to the Ising model. Preprint, 2002; arXiv:math.CO/0211070.

[8] Boulatov, D., Kazakov, V., One-dimensional string theory with vortices as the upside-down matrix oscillator. *Internat. J. Modern Phys. A* **8** (1993), 809–851.

[9] Bouttier, J., Di Francesco, P., Guitter, E., Census of planar maps: from the one-matrix model solution to a combinatorial proof. *Nuclear Phys. B* **645** (2002), 477–499.

[10] Bouttier, J., Di Francesco, P., Guitter, E., Planar maps as labeled mobiles. *Electron. J. Combin.* **11** (2004), Research Paper 69, 27 pp. (electronic).

[11] Brézin, E., Itzykson, C., Parisi, G., and Zuber, J. B., Planar diagrams. *Comm. Math. Phys.* **59** (1978), 35–51.

[12] Chassaing, P., Schaeffer, G., Random planar lattices and integrated superBrownian excursion. *Probab. Theory Related Fields* **128** (2004), 161–212.

[13] Cori, R., Vauquelin, B., Planar maps are well labeled trees. *Canad. J. Math.* **33** (1981), 1023–1042.

[14] Deift, P., Kriecherbauer, T., McLaughlin, K. T.-R., New results on the equilibrium measure for logarithmic potentials in the presence of an external field. *J. Approx. Theory* **95** (1998), 388–475.

[15] Di Francesco, P., and Ginsparg, P., and Zinn-Justin, J., 2D gravity and random matrices. *Phys. Rep.* **254** (1995), 133pp.

[16] Eynard, B., Master loop equations, free energy and correlations for the chain of matrices. *J. High Energy Phys. A* **11** (2003), 018, 45 pp. (electronic).

[17] Ercolani, N. M., and McLaughlin, K. D. T.-R., Asymptotics of the partition function for random matrices via Riemann-Hilbert techniques and applications to graphical enumeration. *Internat. Math. Res. Notices* **14** (2003), 755–820.

[18] Giménez, O., Noy, M., Asymptotic enumeration and limit laws of planar graphs. Preprint, 2005; ariv:math.CO/0501269.

[19] Goulden, I. P., Jackson, D. M., *Combinatorial enumeration*. Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Inc., New York 1983.

[20] Guionnet, A., First order asymptotics of matrix integrals; a rigorous approach towards the understanding of matrix models. *Comm. Math. Phys.* **244** (2004), 527–569.

[21] Guionnet, A., Large deviations and stochastic calculus for large random matrices. *Probab. Surv.* **1** (2004), 72–172.

[22] Guionnet, A., Maurel-Segala, E., Combinatorial aspects of matrix models. *Alea*, to appear 2006 (electronic).

[23] Guionnet, A., Maurel-Segala, E., Second order asymptotics for matrix models. Preprint, 2006; http://front.math.ucdavis.edu/math.PR/0601040.

[24] Guionnet, A., Maïda, M., Character expansion method for the first order asymptotics of a matrix integral. *Probab. Theory Related Fields* **132** (2005), 539–578.

[25] Guionnet, A., Zeitouni, O., Large deviations asymptotics for spherical integrals. *J. Funct. Anal.* **188** (2002), 461–515.

[26] Harary, F., Unsolved problems in the enumeration of graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 63–95.

[27] Kenyon, R., Okounkov, A., Limit shapes and the complex Burgers equation. Preprint, 2005; http://front.math.ucdavis.edu/math.AG/0512573.

[28] Knizhik, V., Polyakov, A., and Zamolodchikov, A., Fractal structure of 2D-quantum gravity. *Mod. Phys. Lett A* **3** (1988), 819–826.

[29] Maurel-Segala, Edouard, High order expansion for matrix models. Preprint, 2006.

[30] Mehta, M.L., *Random matrices*. Third edition, Pure Appl. Math. (Amsterdam) 142, Elsevier/Academic Press, Amsterdam 2004.

[31] Mehta, M. L., A method of integration over matrix variables. *Comm. Math. Phys.* **79** (1981), 327–340.

[32] Schaeffer, G., Bijective census and random generation of Eulerian planar maps with prescribed vertex degrees. *Electron. J. Combin.* **4** (1997), Research Paper 20, 14 pp. (electronic).

[33] 't Hooft, G., Magnetic monopoles in unified gauge theories. *Nuclear Phys. B* **79** (1974), 276–284.

[34] Tutte, W. T., A census of planar triangulations. *Canad. J. Math.* **14** (1962), 21–38.

[35] Tutte, W. T., A new branch of enumerative graph theory. *Bull. Amer. Math. Soc.* **68** (1962), 500–504.

[36] Zvonkin, A., Matrix integrals and map enumeration: an accessible introduction. *Math. Comput. Modelling* **26** (1997), 281–304.

Unité de Mathématiques Pures et Appliquées, CNRS UMR 5669, École Normale Supérieure de Lyon, 46, allée d'Italie, 69364 Lyon Cedex 07, France

E-mail: Alice.GUIONNET@umpa.ens-lyon.fr

# The weak/strong survival transition on trees and nonamenable graphs

Steven P. Lalley*

**Abstract.** Various stochastic processes on nonamenable graphs and manifolds of exponential volume growth exhibit phases that do not occur in the corresponding processes on amenable graphs. Examples include: (1) *branching diffusion* and random walk on hyperbolic space, which for intermediate branching rates may survive globally but not locally; (2) *contact processes* on homogeneous trees, which likewise can survive globally while dying out locally; and (3) *percolation* on Cayley graphs of nonamenable groups, where for certain parameter values infinitely many infinite percolation clusters may coincide. This article surveys some of what is known about the intermediate phases and the upper phase transitions for these processes.

**Mathematics Subject Classification (2000).** Primary 00A05; Secondary 00B10.

**Keywords.** Weak survival, nonamenable graph, contact process, percolation, branching Brownian motion, hyperbolic plane, Fuchsian group.

## 1. Branching Brownian motion and random walk

**B.B.M. in the hyperbolic plane.** Branching Brownian motion in the hyperbolic plane $\mathbb{H}$ is perhaps the simplest process exhibiting the weak/strong survival transition. It evolves as follows: At time 0, a single particle located at a specific point $x_0 \in \mathbb{H}$ begins a Brownian motion. At random exponentially distributed times with mean $1/\lambda$, independent of the motion, the particle undergoes *binary fission*, in which a replicate particle is created at the current location of the fissioning particle. The offspring particles behave as their parents, executing Brownian motions from the places of their births and undergoing further binary fissions at exponentially distributed random times; their behavior is completely independent of their parents' and other particles' behaviors, except for the locations of their births.

The behavior of the branching Brownian motion is controlled by the fission parameter $\lambda$. The size $N_t$ of the population at times $t$ is a simple continuous-time Galton–Watson process with $EN_t = e^{\lambda t}$. If a particle is chosen at random from the $N_t$ particles in existence at time $t$, the distribution of its position has as its density the heat kernel $p_t(x, \cdot)$. This is known to behave asymptotically as $t \to \infty$ like

$$p_t(x, y) \sim C_{x,y} t^{-3/2} \exp\{-t/8\} \tag{1}$$

*Research supported by NSF Grant DMS-0405102.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

for constants $0 < C_{x,y} < \infty$ varying smoothly with $x, y$. Thus, the mean number of particles located in a bounded neighborhood $U$ of $y$ grows/decays roughly as $\exp\{(\lambda - 1/8)t\}$. It is not difficult to deduce the following.

**Proposition 1.** *For $\lambda \le 1/8$, branching Brownian motion survives weakly, that is, for every bounded region $U$ the number of particles located in $U$ is eventually 0, w.p.1. For $\lambda > 1/8$ it survives strongly, that is, for every open set $U$ the number of particles located in $U$ converges to $\infty$ as $t \to \infty$, w.p.1.*

In the weak survival phase, all particle trajectories tend to the boundary circle $\partial\mathbb{H}$ of the hyperbolic plane. Define $\Lambda$ to be the set of all accumulation points of particle trajectories in $\partial\mathbb{H}$. With probability 1, $\Lambda$ is a nonempty, compact subset of $\partial\mathbb{H}$.

**Theorem 2** ([14]). *For $\lambda \in (0, 1/8]$, the limit set $\Lambda$ is, with probability 1, a Cantor set of Hausdorff dimension*

$$\delta = \delta(\lambda) = \frac{1}{2}(1 - \sqrt{1 - 8\lambda}). \tag{2}$$

Observe that as $\lambda \to 1/8$ from below, the Hausdorff dimension approaches $1/2$, not 1 (the dimension of the ambient boundary $\partial\mathbb{H}$), as one might at first suspect. In the strong survival phase $\lambda > 1/8$ the limit set $\Lambda$ is the entire boundary $\partial\mathbb{H}$, so Theorem 2 shows that the Hausdorff dimension behaves discontinuously at the critical parameter $\lambda = 1/8$. Moreover, it shows that the *critical exponent* for $\delta(\lambda)$ at the transition is $1/2$. This, as it turns out, is closely related to the exponent $3/2$ appearing in the asymptotic formula (1) for the heat kernel.

Theorem 2 has been generalized to branching Brownian motion and certain other isotropic branching random walks on higher-dimensional hyperbolic spaces $\mathbb{H}^d$ by Karpelevich, Pechersky, and Suhov [8]: they prove that, for branching Brownian motion in $\mathbb{H}^d$, the Hausdorff dimension $\delta(\lambda)$ of the limit set $\Lambda$ converges up to $(d-1)/2$ as $\lambda$ approaches the critical point from below.

**Branching random walk on $\mathbb{T}^d$.** The existence of a weak survival phase for B.B.M. in $\mathbb{H}$ is a consequence of the exponential decay (1) of the heat kernel. A fundamental theorem of Kesten [9], [10] asserts that exponential decay of random walk transition probabilities is characteristic of nonamenable groups. Thus, branching random walk in any nonamenable group must also have a weak survival phase.

The transition from weak survival to strong survival is understood only for branching random walk on the homogeneous tree $\mathbb{T}^d$ of degree $d \ge 3$ (the Cayley graph of the free product $\Gamma^d := (\mathbb{Z}_2)^{*d}$). Let $\{p_i\}_{i \in A \cup \{1\}}$ be a positive probability distribution on $A \cup \{1\}$ where $A$ is the set of generators of $\Gamma^d$, and denote by $p_n(x, y)$ the $n$-step transition probabilities of the random walk with step distribution $\{p_i\}$. The branching random walk associated with the probability distribution $\{p_i\}$ is constructed as follows: At time $n = 0$, the process is initiated by a single particle located at the site 1 (the root of the tree). The population $X_{n+1}$ of each subsequent generation $n + 1$

is obtained from $X_n$ in two steps: First, each particle $\zeta$ in $X_n$ reproduces, creating a random number $N_\zeta \geq 1$ of replica particles, all located at the same vertex of $\mathbb{T}^d$ as $\zeta$. The distribution of the offspring count $N_\zeta$ is geometric+1 with mean $\lambda > 1$. Second, each particle moves to a randomly chosen neighboring vertex, according to the distribution $\{p_i\}$.

Let $R > 1$ be the spectral radius of the base random walk, that is,

$$R^{-1} := \lim_{n\to\infty} p_n(x, y)^{1/n}. \tag{3}$$

Just as for B.B.M. in $\mathbb{H}$, if the mean offspring number $\lambda$ exceeds $R$ then the number of particles located at the root vertex 1 will explode almost surely. However, if $\lambda \leq R$, then the branching random walk survives only weakly: although the total population size grows exponentially at rate $\lambda$, the number of particles located at any particular vertex will eventually be zero, w.p.1. (That the B.R.W. survives only weakly at the critical point $\lambda = R$ follows because the base random walk is $R$-transient: see [24].) Thus, for $\lambda \leq R$, particle trajectories converge to the space $\partial\mathbb{T}^d$ of ends of the tree. Hence, we may define

$$\Lambda := \{\text{ends in which the BRW survives}\}. \tag{4}$$

The Hausdorff dimension $\delta_H(\Lambda)$, computed with respect to the natural metric[1] on $\partial\mathbb{T}^d$, is a natural measure of the growth of the B.R.W. in the weak survival phase, for the following reason: If $M_m$ is the number of vertices of $\mathbb{T}^d$ at distance $m$ from the root that are ever visited by particles of the B.R.W., then with probability one,

$$\lim_{m\to\infty} M_m^{1/m} = \theta(\lambda) \tag{5}$$

where [7],

$$\delta_H(\Lambda) = \frac{\log\theta(\lambda)}{\log 2}. \tag{6}$$

Denote by $G_x(\lambda)$ the Green's function and $F_x(\lambda)$ the first-passage generating function of the base random walk, that is,

$$\begin{aligned} G_x(\lambda) &= \sum_{n=0}^{\infty} p_n(1, x), \\ F_x(\lambda) &= G_x(\lambda)/G_1(\lambda). \end{aligned} \tag{7}$$

**Theorem 3** ([7]). *The Malthusian parameter $\theta(\lambda)$ is the unique positive number such that*

$$\sum_{i\in A} \frac{F_i(\lambda)}{F_i(\lambda) + \theta(\lambda)} = 1. \tag{8}$$

[1]The natural metric $d$ is defined by $d(\alpha, \beta) = 2^{-N(\alpha,\beta)}$, where $N(\alpha, \beta)$ denotes the number of common edges in the geodesic segments from the root to $\alpha$ and from the root to $\beta$.

*This parameter has critical exponent* $1/2$ *at the critical point* $\lambda = R$*: that is, there exists a constant* $C > 0$ *such that as* $\lambda \to R$ *from below,*

$$\theta(R) - \theta(\lambda) \sim C\sqrt{R - \lambda}. \tag{9}$$

*Furthermore,*

$$\theta(\lambda) \le \sqrt{d-1} \tag{10}$$

*and equality holds if and only if the step distribution* $\{p_i\}$ *is isotropic.*

The formula (8) makes it clear that the critical exponent $1/2$ in (9) is related to the exponent $3/2$ occurring in the power law

$$p_n(1, x) \sim C_x R^{-n} n^{-3/2} \tag{11}$$

for the base random walk transition probabilities [24]. This is because (11) is determined by the singularity of the Green's function $G_x(\lambda)$ at $\lambda = R$, by standard Tauberian theorems, and this in turn has the same singular asymptotics as the first-passage generating functions $F_x(\lambda)$. It is conjectured that the local limit theorem (11) holds more generally for random walks on nonelementary Fuchsian groups (discrete groups of isometries of the hyperbolic plane), but this has been proved only for Fuchsian groups containing free groups as subgroups of finite index. It is natural to expect that branching random walks on such groups will have weak/strong survival transitions of the same type as on homogeneous trees.

## 2. Contact processes

**Weak survival.** Let $G = (V, E)$ be the Cayley graph of a finitely generated group $\Gamma$, with edges labeled by elements of the generating set $A$, and let $\mathcal{P} = \{p_a\}_{a \in A}$ be a probability distribution on $A$. The *contact process* with intensity parameter $\lambda > 0$ and infection rates $\mathcal{P}$ is a Markov process on the configuration space $\{0, 1\}^V$ (here $0 =$ "healthy" and $1 =$ "infected") that evolves as follows: (A) Infected vertices "recover" (become healthy) at rate 1. (B) Healthy vertices $x$ become infected at rate

$$\lambda \sum_{j \in A : xj \in \xi_t} p_j$$

where $\xi_t$ is the set of vertices that are infected at time $t$. If the probability distribution $\mathcal{P}$ is uniform on $A$, the contact process is said to be *isotropic*. If the generating set $A$ and the rates $\mathcal{P}$ are symmetric (that is, $a \in A$ implies $a^{-1} \in A$ and $p_a = p_{a^{-1}}$) then the contact process is said to be *symmetric*. The default initial condition is $\xi_0 = \{1\}$, where 1 denotes the group identity. The contact process is said to survive weakly if $\xi_t \neq \emptyset$ for all $t > 0$ but $\xi_t \cap F = \emptyset$ eventually for every finite set $F \subset V$. It survives strongly if for every nonempty set $F \subset V$, the intersection $\xi_t \cap F$ is nonempty at indefinitely large times $t$.

Most of what is known about existence of weak and strong survival phases for contact processes is restricted to groups $\Gamma$ whose associated Cayley graphs $G$ are trees.

**Theorem 4** ([20], [18], [23], [25]). *Assume that $G$ is the homogeneous tree $\mathbb{T}^d$ of degree $d \geq 3$. Then for the isotropic contact process there exist constants $0 < \lambda_c < \lambda_u < \infty$ (depending on $d$) so that*

(a) $\lambda \leq \lambda_c \Longrightarrow$ *ultimate extinction with probability* 1.

(b) $\lambda_c < \lambda \leq \lambda_u \Longrightarrow$ *weak survival with positive probability.*

(c) $\lambda > \lambda_u \Longrightarrow$ *strong survival with positive probability.*

Unlike the corresponding results for branching random walks and branching Brownian motion on hyperbolic spaces, Theorem 4 is surprisingly difficult (at least for small $d$). The proofs in [20], [18], and [23] rely heavily on both isotropy and the absence of cycles in the graph. The following result weakens the isotropy requirement.

**Theorem 5** ([16]). *Assume that $G$ is a homogeneous tree of degree $d \geq 3$. Assume further that the rates $\mathcal{P}$ are symmetric and* weakly isotropic *in the sense that there are generators $a, b$ with $a \neq b^{\pm 1}$ so that $p_a = p_b$. Then the contact process with rates $\mathcal{P}$ has a weak survival phase, that is, there exist $0 < \lambda_c < \lambda_u < \infty$ so that conclusions* (a)*,* (b)*, and* (c) *of Theorem* 4 *are valid.*

Alan Stacey (unpublished) has recently shown that the weak isotropy hypothesis is unnecessary.

**Size of the limit set.** Consider now the isotropic contact process on a homogeneous tree $\mathbb{T}^d$ of degree $d \geq 3$. By Theorem 4, there is a nontrivial weak survival phase $\lambda_c < \lambda < \lambda_u$. By definition of weak survival, any finite set of vertices must eventually be vacated, with probability 1. Therefore, the set of occupied vertices must recede to $\partial\mathbb{T}^d$ as $t \to \infty$. As for branching random walk in $\mathbb{T}^d$, define the *limit set* $\Lambda$ to be the set of ends in which the contact process survives.

**Theorem 6** ([15]). *For the isotropic contact process in the weak survival phase, the Hausdorff dimension $\delta_H(\Lambda)$ of the limit set $\Lambda$ is a.s. constant on the event of survival, and satisfies the inequality*

$$\delta_H(\Lambda) \leq \frac{1}{2}\delta_H(\partial\mathbb{T}^d). \tag{12}$$

*Furthermore, $\delta_H(\Lambda)$ is continuous* [22] *and strictly increasing* [12] *in the parameter $\lambda$, and equality in* (12) *holds at $\lambda = \lambda_u$.*

The inequality (12) holds for essentially the same reason as for branching random walk: If $\delta_H(\Lambda)$ were greater than $(1/2)\delta_H(\partial\mathbb{T}^d)$ then there would be particle trajectories extending from the root vertex to vertices far from the root and then back to

the root, contradicting weak survival. As for branching random walk on a homogeneous tree, the inequality (12) remains valid for nonisotropic but *symmetric* contact processes.

The Hausdorff dimension of the limit set $\Lambda$ is simply related to a hitting probability associated to the contact process. Let $x \in V$ be a vertex at distance $n$ from the root vertex 1, and define

$$u_n = P\{x \in \xi_t \text{ for some } t > 0\} \tag{13}$$

to be the probability that the vertex $x$ is ever infected. By isotropy, this probability is the same for all vertices $x$ at distance $n$ from 1. It is apparent that $u_{m+n} \geq u_m u_n$, and so

$$\lim_{n\to\infty} u_n^{1/n} := \beta = \beta(\lambda) \tag{14}$$

exists and is $\leq 1$.

**Theorem 7** ([15]).

$$\delta_H(\Lambda) = -\frac{\log(d-1)\beta}{\log 2}. \tag{15}$$

**Critical exponent.** For the isotropic contact process on the tree $\mathbb{T}^d$ of degree $d \geq 3$, the Hausdorff dimension $\delta(\lambda) := \delta_H(\Lambda)$ varies continuously with the intensity $\lambda$ for $\lambda \leq \lambda_u$ [22], and increases *strictly* with $\lambda$ in the interval $\{\lambda : \delta(\lambda) < 1/2\delta_H(\partial\mathbb{T}^d)\}$ [12]. Define

$$\lambda_* = \sup\{\lambda : \delta(\lambda) < 1/2\delta_H(\partial\mathbb{T}^d)\}. \tag{16}$$

**Conjecture 8.** $\lambda_* = \lambda_u$.

Recall that the critical exponent at the upper critical point for branching random walk on $\mathbb{T}^d$ is $1/2$, by formula (9). It is believed that the phase transition for the contact process is of the same type as the corresponding phase transition for branching random walk, and so it is natural to conjecture that the critical exponent is again $1/2$:

**Conjecture 9.**

$$\lim_{\lambda\to\lambda_*} \frac{\log(\delta_H(\partial\mathbb{T}^d) - 2\delta(\lambda))}{\lambda_* - \lambda} = \frac{1}{2}.$$

Further evidence for the truth of this conjecture is provided by the main result of [17], which we now describe. Consider the isotropic contact process $\xi_t$ on $\mathbb{T}^d$ in the weak survival phase $\lambda \in (\lambda_c, \lambda_u]$. For any site $x$, the total infection time

$$J(x) := \int_0^\infty \mathbf{1}\{x \in \xi_t\}\, dt \tag{17}$$

is finite with probability 1. It is known [12] that if $\lambda < \lambda_*$ then $P\{x \in \xi_t\}$ is exponentially decaying in $t$, and so $EJ(x) < \infty$. Because the hitting probability $u_n$ decays exponentially in $n$ even at the critical point $\lambda = \lambda_*$, it is natural to expect that the conditional expectation of $J(x)$ given $J(x) > 0$ is finite.

**Conjecture 10.** There exists a constant $C = C_d$ depending only on the degree $d$ of the tree $\mathbb{T}^d$ such that, for every vertex $x$ and all $\lambda \leq \lambda_*$,

$$E(J(x) \mid J(x) > 0) \leq C. \tag{18}$$

The analogous statement is known to be true for branching random walk. For the contact process, it is at least as plausible as Conjecture 9.

**Theorem 11** ([17]). *If Conjecture* 10 *is true then there is a finite constant* $C = C_d$ *so that for all* $\lambda < \lambda_*$,

$$\frac{1}{2}\delta_H(\partial\mathbb{T}^d) - \delta(\lambda) \leq C\sqrt{\lambda_* - \lambda}. \tag{19}$$

Thus, if Conjecture 10 is true, and if there is a critical exponent, then it cannot be less than $1/2$. The proof of Theorem 11 in [17] also suggests that $1/2$ is the correct value, as the inequalities in the proof are very likely approximate equalities.

## 3. Percolation

**Coexistence of infinite clusters.** In *Bernoulli site* (*resp., bond*) *percolation* on a graph $G$, vertices (resp., edges) are colored blue or red independently, blue with probability $p$, red with probability $1 - p$. For brevity we shall discuss only site percolation; however, most of the results and conjectures have natural analogues for bond percolation.

In site percolation, interest focuses on the connected clusters of blue vertices, and in particular on the existence/uniqueness and geometry of infinite blue clusters. *Percolation* is said to occur if there is an infinite blue cluster. For any infinite graph there exists a unique threshold $p_c \in [0, 1]$ for the Bernoulli parameter $p$ above which percolation occurs with positive probability, and below which it occurs with probability zero. Burton and Keane [5] showed that if the ambient graph $G$ is the Cayley graph of a finitely generated, amenable group, then infinite blue clusters, if they exist, are unique w.p.1. Grimmett and Newman [6] showed that uniqueness of infinite clusters need not hold in nonamenable graphs: in particular, they showed that Bernoulli percolation on $\mathbb{Z} \times \mathbb{T}^d$ has infinitely many infinite clusters for certain values of $p$, provided the degree $d$ is sufficiently large.

A graph is called *transitive* if its automorphism group acts transitively on the vertex set, and is called *nonamenable* if there exists a constant $\varepsilon > 0$ such that for any finite set $V_0$ of vertices, $|\partial V_0| > \varepsilon |V_0|$. (Here $\partial V_0$ denotes the set of vertices not in $V_0$ that are connected by edges to vertices of $V_0$.)

**Conjecture 12** ([1]). If $G$ is a transitive, nonamenable graph, then there exists a nonempty interval $I = (p_c, p_u)$ such that for all $p \in I$, Bernoulli-$p$ site percolation has infinitely many infinite blue clusters.

In full generality, this remains unresolved. However, two important results have been obtained:

**Theorem 13** ([2]). *Let $G$ be a transitive, nonamenable, planar graph with one end. Then there exist constants $0 < p_c < p_u < 1$ such that*

(a) $p \le p_c \Longrightarrow$ *no infinite blue clusters;*

(b) $p_c < p < p_u \Longrightarrow$ *infinitely many infinite blue clusters;*

(c) $p_u \le p \Longrightarrow$ *one infinite blue cluster.*

A connected, transitive graph is said to have one end if the subgraph obtained by deleting any finite set of vertices remains connected. Observe that the theorem asserts uniqueness of the infinite cluster at the upper transition point $p_u$: this contrasts with the analogous transition on $\mathbb{Z} \times \mathbb{T}^d$, where it is known [21] that at least for large $d$ there are infinitely many infinite clusters at $p_u$. See [3], [4] fr discussion of related issues.

**Theorem 14** ([19]). *For every finitely generated, nonamenable group there is a Cayley graph for which Conjecture* 12 *is true.*

**Percolation clusters in hyperbolic tessellations.** A *Fuchsian group* is a discrete group of isometries of the hyperbolic plane $\mathbb{H}$. Let $\Gamma$ be a co-compact Fuchsian group with finite, symmetric generating set $A$, and let $G$ be the Cayley graph. Then $G$ may be naturally embedded in $\mathbb{H}$ in such a way that edges are geodesic segments in $\mathbb{H}$, and so that any compact subset of $\mathbb{H}$ contains only finitely many vertices of $G$. Denote by $x_0 \in \mathbb{H}$ the vertex of $G$ corresponding to the group identity $1 \in \Gamma$.

Consider Bernoulli-$p$ site percolation on $G$, and let $K$ be the connected blue cluster containing $x_0$. On the event that $K$ is infinite, vertices in $K$ will accumulate at the boundary circle $\partial\mathbb{H}$: define $\Lambda$ to be the (closed) set of accumulation points.

**Theorem 15** ([11]). *If $p_c < p < p_u$ then on the event that $K$ is infinite the limit set $\Lambda$ is a Cantor set of Lebesgue measure* 0.

Recall that for the contact process on a homogeneous tree, the Hausdorff dimension of the limit set is discontinuous at the transition from weak to strong survival. It is natural to ask how the Hausdorff dimension of the limit set $\Lambda$ of a percolation cluster in $\mathbb{H}$ behaves as $p \to p_u$ from below.

**Theorem 16** ([13]). *For each $p \in (p_c, p_u)$ the Hausdorff dimension $\delta(p)$ of the limit set $\Lambda$ in Bernoulli-$p$ site percolation is almost surely constant. The function $p \mapsto \delta(p)$ is continuous and strictly increasing in $p$, with limit* 1 *as $p \to p_u$.*

Thus, the nature of the phase transition for Bernoulli percolation seems to be different from that for contact processes and branching random walks.

The use of $\delta(p)$ as a measure of the size of percolation clusters is not unreasonable for Fuchsian hyperbolic groups, but a more natural measure might be the volume growth rate

$$\varrho = \varrho(p) = \lim_{R\to\infty} R^{-1} \log \operatorname{card} K_R \tag{20}$$

where $K_R$ denotes the intersection of the cluster $K$ with the ball of radius $R$ centered at the root $x_0$. (For Fuchsian groups, it can be shown that $\varrho(p) = \delta(p)$.) Whereas $\delta(p)$ depends for its definition on the existence of a geometric boundary of the ambient space, $\varrho(p)$ can in principle be used for an arbitrary infinite group, using the graph metric to measure volume growth (existence of the limit must be proved, of course).

**Conjecture 17.** Assume that $G$ is a transitive, nonamenable graph with nonempty coexistence phase $(p_c, p_u)$. If there is a unique infinite cluster a.s. at $p = p_u$, then $\varrho(p)$ converges, as $p \to p_u$ from below, to the volume growth rate of the ambient graph $G$.

The proof of Theorem 16 leads to an interesting variational formula for the Hausdorff dimension $\delta(p)$ of percolation clusters in co-compact Fuchsian groups. Define the *connectivity function* $\tau : \Gamma \to [0, 1]$ as follows:

$$\tau(x) = \tau(x; p) = P_p\{x \in K\}. \tag{21}$$

By the FKG inequality, $\tau$ satisfies a log-subadditivity inequality on $\Gamma$:

$$\tau(xy) \geq \tau(x)\tau(y). \tag{22}$$

The function $\tau$ may be extended to a function on the entire hyperbolic plane $\mathbb{H}$ by setting $\tau(w) = \tau(x)$ where $x$ is the vertex of the Cayley graph $G$ nearest $w$ (with some convention for ties). Now consider the geodesic flow $\Phi_t$ on the unit tangent bundle of $\mathbb{H}/\Gamma$: geodesics may be lifted to $\mathbb{H}$, and so by (21) the log-connectivity function evaluated along geodesics is subadditive for the geodesic flow. Therefore, by Kingman's subadditive ergodic theorem, for any ergodic, invariant probability measure $\mu$ for the flow $\Phi_t$, there exists a constant $\beta(\mu) = \beta(\mu; p)$ so that $\mu$-almost surely, the connectivity function $\tau$ decays at rate $\beta(\mu)$ along geodesics, that is,

$$\lim_{t\to\infty} t^{-1} \log \tau \circ \Phi_t = \beta(\mu). \tag{23}$$

Denote by $\mathcal{I}$ the set of all invariant probability measures for the geodesic flow, and by $h(\mu)$ the Kolmogorov–Sinai entropy of the geodesic flow relative to the invariant measure $\mu$.

**Theorem 18** ([13]). *The decay rate function is jointly continuous in $\mu$ and $p$, and for each $\mu$ is strictly increasing in $p$ for $p \in (p_c, p_u)$. Moreover, $\beta(\mu; p) = 0$ for all $\mu \in \mathcal{I}$ and all $p \geq p_u$. For every $p$ the Hausdorff dimension $\delta(p)$ of the limit set $\Lambda$ is, $P_p$-almost surely on the event $|K| = \infty$, given by*

$$\delta(p) = \max_{\mu\in\mathcal{I}} (h(\mu) + \beta(\mu; p)). \tag{24}$$

Although it is by no means obvious, this is the natural analogue of formula (8) for the H.D. of the limit set of a branching random walk on $\mathbb{T}^d$. Formula (24) has further implications for the geometry of percolation clusters. Recall that $K_R$ is the intersection of the percolation cluster $K$ with the (hyperbolic) ball of radius $R$ centered at $x_0$. The cardinality of $K_R$ grows exponentially, at rate $\delta(p)$ (see (20)), and most of the vertices in $K_R$ are at distance nearly $R$ from $x_0$. Suppose that one of these is chosen at random: then the geodesic ray from $x_0$ through the randomly chosen vertex of $K_R$ will be approximately $\mu$-generic, where $\mu$ is the maximizing measure in (24).

## References


[1] Benjamini, I., and Schramm, O., Percolation beyond $\mathbb{Z}^d$, many questions and a few answers. *Electron. Comm. Probab.* **1** (1996), 71–82.

[2] Benjamini, I., and Schramm, O., Percolation in the hyperbolic plane. *J. Amer. Math. Soc.* **14** (2001), 487–507.

[3] Benjamini, I., Lyons, R., Peres, Y., and Schramm, O., Critical percolation on any nonamenable group has no infinite clusters. *Ann. Probab.* **27** (1998), 1347–1356.

[4] Benjamini, I., Lyons, R., Peres, Y., and Schramm, O., Group invariant percolation. *Geom. Funct. Anal.* **9** (1999), 29–66.

[5] Burton, R., and Keane, M., Density and uniqueness in percolation. *Comm. Math. Phys.* **121** (1989), 501–505.

[6] Grimmett, G., and Newman, C., Percolation in $\infty + 1$ dimensions. In *Disorder in physical systems*, Oxford Sci. Publ., Oxford University Press, New York 1990, 167–190.

[7] Hueter, I., and Lalley, S., Anisotropic branching random walks on homogeneous trees. *Probab. Theory Related Fields* **116** (2000), 57–88.

[8] Karpelevich, F., Pechersky, E., and Suhov, Y., A phase transition for hyperbolic branching processes. *Comm. Math. Phys.* **195** (1998), 627–642.

[9] Kesten, H., Full Banach mean values on countable groups. *Math. Scand.* **7** (1959), 146–156.

[10] Kesten, H., Symmetric random walks on groups. *Trans. Amer. Math. Soc.* **92** (1959), 336–354.

[11] Lalley, S., Percolation on Fuchsian groups. *Ann. Inst. H. Poincaré Probab. Statist.* **34** (1998), 151–177.

[12] Lalley, S., Growth profile and invariant measures for the contact process on a homogeneous tree. *Ann. Probab.* **27**, 206–225; Correction: ibid. **30** (1999), 2108–2112.

[13] Lalley, S., Percolation clusters in hyperbolic tessellations. *Geom. Funct. Anal.* **11** (2001), 971–1030.

[14] Lalley, S., and Sellke, T., Hyperbolic branching Brownian motion. *Probab. Theory Related Fields* **108** (1997), 171–192.

[15] Lalley, S., and Sellke, T., Limit set of a weakly supercritical contact process on a homogeneous tree. *Ann. Probab.* **26** (1998), 644–657.

[16] Lalley, S., and Sellke, T., Anisotropic contact processes on homogeneous trees. *Stochastic Process. Appl.* **101** (2002), 163–183.

[17] Lalley, S., and Sellke, T., The weak survival/strong survival phase transition for the contact process on a homogeneous tree. *Bull. Braz. Math. Soc.* (*N.S.*) **33** (2002), 341–350.

[18] Liggett, T., Multiple transition points for the contact process on the binary tree. *Ann. Probab.* **24** (1996), 1675–1710.

[19] Pak, I., and Smirnova-Nagnibeda, T., On non-uniqueness of percolation on nonamenable Cayley graphs. *C. R. Acad. Sci. Paris Ser. I Math.* **330** (2000), 495–500.

[20] Pemantle, R., The contact process on trees. *Ann. Probab.* **20** (1992), 2089–2169.

[21] Schonmann, R., Percolation in $\infty+1$ dimensions at the uniqueness threshold. In *Perplexing problems in probability*, Progr. Probab. 44, Birkhäuser, Boston, MA, 1999, 53–67.

[22] Schonmann, R., The triangle condition for contact processes on homogeneous trees. *J. Statist. Phys.* **90** (1999), 1429–1440.

[23] Stacey, A., The existence of an intermediate phase for the contact process on trees. *Ann. Probab.* **24** (1996), 1711–1726.

[24] Woess, W., *Random walks on infinite graphs and groups.* Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

[25] Zhang, Y., The complete convergence theorem of the contact process on trees. *Ann. Probab.* **24** (1996), 1408–1443.

Department of Statistics, University of Chicago, 5734 University Avenue, Chicago IL 60637, U.S.A.

E-mail: lalley@galton.uchicago.edu

# New developments in stochastic dynamics

Yves Le Jan

**Abstract.** Flows of random coalescing maps or flows of random transition probabilities can arise from simple stochastic differential equations when Ito's theory of strong solutions ceases to apply.

**Mathematics Subject Classification (2000).** 60H10, 60H40, 60G51, 76F05.

**Keywords.** Stochastic differential equations, stochastic flows, coalescing flows, noise.

## Introduction

A stationary motion on the real line with independent increments is described by a Levy process, or equivalently by a convolution semigroup of probability measures. This naturally extends to "rigid" motions represented by Levy processes on Lie groups. If one assumes the continuity of the paths, a convolution semigroup on a Lie group $G$ is determined by an element of the Lie algebra $\mathfrak{g}$ (the drift) and a scalar product on $\mathfrak{g}$ (the diffusion matrix). We call them the local characteristics of the convolution semigroup. We will be interested in stationary "fluid" random evolutions which have independent increments. They can be modelled by stochastic differential equations driven by Wiener processes. These have been studied for more than fifty years. In particular, it was shown that stochastic flows driven by smooth Brownian vector fields on a compact manifold define flows of diffeomorphisms. This is also true on non compact manifolds under appropriate conditions of non explosion ([7]). Such flows can be viewed as infinitely divisible limits of products of i.i.d. (independant and indentically distributed) random diffeomorphisms, and the theory is at least formally very similar to the theory of Brownian motion on Lie groups ([14], [19], [12]). Their laws can be viewed as convolution semigroups of probability measures on the group of diffeomorphisms. They are characterized by two functions: the covariance of the Brownian vector field, or equivalently its auto reproducing space, which plays the role of a metric on the Lie algebra and a drift vector field. One can similarly extend the notion of Levy processes by introducing Poisson measures on the group itself. A remarkable result of Tsirelson (cf. [24]) shows that there is essentially no other way to define a process $X_t$ with stationary independent increments on the unitary group of the Hilbert space. The *noise* defined by the increments of the flow, i.e. the family of $\sigma$-fields $\mathcal{F}_{s,t} = \sigma\{X_v X_u^{-1}, s \leq u \leq v \leq t\}$ is classical, i.e. generated by additive

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

increments of Wiener and (or) Poisson processes. But in recent years it appeared that this picture was not complete: indeed flows of non invertible transformations, and as well of transition probabilities, appear to play an important role in the theory.

To be more precise, on a compact manifold let $V_0, V_1, \dots, V_n$ be vector fields and $B^1, \dots, B^n$ be independent Brownian motions. Consider the SDE

$$dX_t = \sum_{k=1}^{n} V_k(X_t) \circ dB_t^k + V_0(X_t)\, dt, \tag{1}$$

which equivalently can be written as

$$df(X_t) = \sum_{k=1}^{n} V_k f(X_t) dB_t^k + \frac{1}{2} Af(X_t) dt \tag{2}$$

for every smooth function $f$ and $Af = \sum_{k=1}^{n} V_k(V_k f) + V_0 f$. Observe that $Af^2 - 2fAf = \sum_{k=1}^{n} (V_k f)^2$. Then strong solutions of this SDE produce a flow of maps $\varphi_t$ such that, for every $x$, $\varphi_t(x)$ is a strong solution of the SDE with $\varphi_0(x) = x$. When the vector fields are smooth, strong solutions are known to exist and to be unique. The framework can be extended to include flows of maps driven by vector field valued Brownian motions, which means essentially that $n = \infty$ (see for example [3], [12], [14], [19], [23]).

In a joint paper ([15]) with Olivier Raimond this was extended again to include flows of Markovian operators $S_t$ which are solutions of the SPDE

$$dS_t f = \sum_{k=1}^{\infty} S_t(V_k f) dB_t^k + \frac{1}{2} S_t(Af) dt, \tag{3}$$

assuming that the covariance function $C = \sum_{k=1}^{\infty} V_k \otimes V_k$ of the Brownian vector field $\sum_{k=1}^{\infty} V_k B^k$ is compatible with $A$, namely that

$$Af^2 - 2fAf \geq \sum_{k=1}^{\infty} (V_k f)^2. \tag{4}$$

Existence and uniqueness of a flow of Markovian operators $S_t$, which is a Wiener solution of the previous SPDE in the sense that $S_t$ is a function of the Brownian paths $(B^i)_{i \geq 1}$ up to time $t$, hold under rather weak assumptions.

The local characteristics of these flows are given by $A$ and the covariance function $C$, and they determine the SDE or the SPDE. Under Lipschitz conditions we actually get strong solutions of stochastic differential equations. These solutions are of a regular type, namely:

(a) The probability that two points thrown in the fluid at the same time and at distance $\varepsilon$, separate at distance one in one unit of time tends to 0 as $\varepsilon$ tends to 0.

(b) Such points will never hit each other.

But it was shown in [15] that covariance functions which are not smooth on the diagonal (e.g. covariance associated with Sobolev norms of order between $d/2$ and $(d+2)/2$, $d$ being the dimension of the space) can produce Wiener solutions which define random evolutions of different types:

– turbulent evolutions where (a) is not satisfied, which means that two points thrown initially at the same place separate, even when there is no pure diffusion, i.e. that $Af^2 - 2fAf = \sum_{k=1}^{\infty}(V_k f)^2$;

– coalescing evolutions where (b) does not hold.

That paper was motivated by the works of physicists working on the Kraichnan model for turbulent advection (cf., for example, [9], [10], [4], [6]).

In a subsequent paper [16] we adopted a more general approach based on consistent systems of $n$-point Markovian Feller semigroups which can be viewed as determining the law of the motion of $n$ indivisible points thrown into the fluid. Regular and coalescing evolutions are represented by flows of maps, and turbulent evolutions by flows of probability kernels $K_{s,t}(x, dy)$ describing how a point mass (made of a continuum of indivisible points) in $x$ at time $s$ is spread at time $t$. (Note that in this case, the motion of an indivisible point is not fully determined by the flow.)

Among turbulent evolutions, we can distinguish the intermediate ones where two points thrown in the fluid at the same place separate but can meet later, i.e. where (a) and (b) are both not satisfied. These flows can always be coupled with a coalescent flow.

Let us explain in more detail the contents of the paper. We give in the first section construction results from [16], which generalize a theorem by De Finetti on exchangeable variables. A stochastic flow of kernels $K$ is associated with a general compatible family $(\mathsf{P}_t^{(n)},\ n \geq 1)$ of Feller semigroups. The flow $K$ is induced by a flow of measurable mappings when

$$\mathsf{P}_t^{(2)} f^{\otimes 2}(x, x) = \mathsf{P}_t f^2(x),$$

for all $f \in C(M)$, $x \in M$ and $t \geq 0$. The Markov process associated with $\mathsf{P}_t^{(n)}$ represents the motion of $n$ indivisible points thrown in the fluid. The key point is that the two notions are shown to be equivalent: the law of a stochastic flow of kernels is uniquely determined by the compatible system of $n$-point motions.

In Section 2 we define the noise associated with a flow and recall the notion of "black noise" introduced by Tsirelson.

Coalescing flows are defined in Section 3. A coalescing flow can be obtained from any flow of kernels the two-point motion of which hits the diagonal. Then the original flow is recovered by filtering the coalescing flow with respect to a sub-noise.

We give the example of Arratia's flow and consider briefly sticky flows.

In Section 4, we present the result of [18] in which the classification of solutions of Tanaka's equation is given.

Finally, in Section 5, we consider stochastic flows on the circle defined by SDE's driven by the white noise $W$, which exhibit most of the features of more general isotropic flows considered in [15] and [16].

## 1. Flows and their construction

This first section is rather formal since we chose to give a precise result. Its intuitive content is rather simple: flows of maps, and more generally flows of transition kernels, are described by their moments which are Markovian semigroups describing the motion of any finite number of points transported by the flow. We refer to [24] for an alternative approach to this construction.

### 1.1. Flows of maps.

Let $M$ be a compact separable metric space.

**Definition 1.1.1.** Let $(\mathsf{P}_t^{(n)},\ n \geq 1)$ be a family of Feller semigroups[1], defined on $M^n$ and acting on $C(M^n)$, respectively. We say that this family is *consistent* as soon as for all $k \leq n$,

$$\mathsf{P}_t^{(k)} f(x_1, \ldots, x_k) = \mathsf{P}_t^{(n)} g(y_1, \ldots, y_n), \tag{1.1}$$

where $f$ and $g$ are any continuous functions such that

$$g(y_1, \ldots, y_n) = f(y_{i_1}, \ldots, y_{i_k}) \tag{1.2}$$

with $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ and $(x_1, \ldots, x_k) = (y_{i_1}, \ldots, y_{i_k})$. We will denote by $\mathsf{P}^{(n)}_{(x_1,\ldots,x_n)}$ the law of the Markov process associated with $\mathsf{P}_t^{(n)}$ starting from $(x_1, \ldots, x_n)$.

This Markov process will be called the $n$-point motion (see also [21]).

We equip $M$ with its Borel $\sigma$-field $\mathcal{B}(M)$. Let $(F, \mathcal{F})$ be the space of measurable mappings on $M$ equipped with the $\sigma$-field generated by the evaluations at $x$ for all $x$ in $M$.

**Definition 1.1.2.** A probability measure $\mathsf{Q}$ on $(F, \mathcal{F})$ is called *regular* if there exists a measurable mapping $\mathcal{J}\colon (F, \mathcal{F}) \to (F, \mathcal{F})$ such that

$$(M \times F, \mathcal{B}(M) \otimes \mathcal{F}) \to (M, \mathcal{B}(M)),$$
$$(x, \varphi) \mapsto \mathcal{J}(\varphi)(x)$$

[1] $\mathsf{P}_t^{(n)}$ is a Feller semigroup on $M^n$ if and only if $\mathsf{P}_t^{(n)}$ is positive (i.e. $\mathsf{P}_t^{(n)} f \geq 0$ for every $f \geq 0$), $\mathsf{P}_t^{(n)} 1 = 1$ and for every continuous function $f$, $\mathsf{P}_t^{(n)} f$ is continuous and $\lim_{t\to 0} \mathsf{P}_t^{(n)} f(x) = f(x)$, which implies the uniform convergence of $\mathsf{P}_t^{(n)} f$ towards $f$ .

is measurable and for every $x \in M$,

$$\mathsf{Q}(d\varphi)\text{-a.s.}, \quad \mathcal{J}(\varphi)(x) = \varphi(x), \tag{1.3}$$

i.e. $\mathcal{J}$ is a measurable modification of the identity mapping on $(F, \mathcal{F}, \mathsf{Q})$. We call it a measurable presentation of $\mathsf{Q}$.

Let $\mathsf{Q}_1$ and $\mathsf{Q}_2$ be two probability measures on $(F, \mathcal{F})$. Assume that $\mathsf{Q}_1$ is regular. Let $\mathcal{J}$ be a measurable presentation of $\mathsf{Q}_1$. Then the mapping

$$\begin{aligned}(F^2, \mathcal{F}^{\otimes 2}) &\to (F, \mathcal{F}),\\ (\varphi_1, \varphi_2) &\mapsto \mathcal{J}(\varphi_1) \circ \varphi_2\end{aligned}$$

is measurable. Moreover, if $\mathcal{J}'$ is another measurable presentation of $\mathsf{Q}_1$, then for every $x \in M$,

$$\mathsf{Q}_1(d\varphi_1) \otimes \mathsf{Q}_2(d\varphi_2)\text{-a.s.}, \quad \mathcal{J}(\varphi_1) \circ \varphi_2(x) = \mathcal{J}'(\varphi_1) \circ \varphi_2(x). \tag{1.4}$$

Note that $(\varphi_1, \varphi_2) \mapsto \mathcal{J}(\varphi_1) \circ \varphi_2$ is measurable, but $(\varphi_1, \varphi_2) \mapsto \varphi_1 \circ \varphi_2$ is not measurable.

**Definition 1.1.3.** The *convolution product* of $\mathsf{Q}_1$ and $\mathsf{Q}_2$, denoted by $\mathsf{Q}_1 * \mathsf{Q}_2$, is the law of the random variable $(\varphi_1, \varphi_2) \mapsto \mathcal{J}(\varphi_1) \circ \varphi_2$ defined on the probability space $(F^2, \mathcal{F}^{\otimes 2}, \mathsf{Q}_1 \otimes \mathsf{Q}_2)$. A *convolution semigroup* on $(F, \mathcal{F})$ is a family $(\mathsf{Q}_t)_{t \geq 0}$ of regular probability measures on $(F, \mathcal{F})$ such that for all nonnegative $s$ and $t$, $\mathsf{Q}_{s+t} = \mathsf{Q}_s * \mathsf{Q}_t$.

A convolution semigroup $(\mathsf{Q}_t)_{t \geq 0}$ on $(F, \mathcal{F})$ is called Feller if

(i) for all $f \in C(M)$, $\lim_{t \to 0} \sup_{x \in M} \int (f \circ \varphi(x) - f(x))^2 \mathsf{Q}_t(d\varphi) = 0$;

(ii) for all $f \in C(M)$ and $t \geq 0$, $\lim_{d(x,y) \to 0} \int (f \circ \varphi(x) - f \circ \varphi(y))^2 \mathsf{Q}_t(d\varphi) = 0$.

Let $(\mathsf{Q}_t)_{t \geq 0}$ be a Feller convolution semigroup on $(F, \mathcal{F})$. For all $n \geq 1$, $f \in C(M^n)$ and $x \in M^n$ set

$$\mathsf{P}_t^{(n)} f(x) = \int f \circ \varphi^{\otimes n}(x)\, \mathsf{Q}_t(d\varphi). \tag{1.5}$$

Then $(\mathsf{P}_t^{(n)},\ n \geq 1)$ is a compatible family of Feller semigroups on $M$ satisfying

$$\mathsf{P}_t^{(2)} f^{\otimes 2}(x, x) = \mathsf{P}_t f^2(x) \tag{1.6}$$

for all $f \in C(M)$, $x \in M$ and $t \geq 0$. *The semigroup* $(\mathsf{Q}_t)_{t \geq 0}$ *is uniquely determined by* $(\mathsf{P}_t^{(n)},\ n \geq 1)$.

In the following we will consider only probability spaces $(\Omega, \mathcal{A}, \mathsf{P})$ which are separable, i.e., the corresponding Hilbert space $L^2(\Omega, \mathcal{A}, \mathsf{P})$ is separable.

**Definition 1.1.4.** Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and $(T_h)_{h\in\mathbb{R}}$ a one parameter group of $\mathsf{P}$-preserving $L^2$-continuous transformations of $\Omega$. A family of $(F, \mathcal{F})$-valued random variables $(\varphi_{s,t},\ s \le t)$ is called a *measurable stochastic flow of mappings* if for all $s \le t$ the mapping

$$(M \times \Omega, \mathcal{B}(M) \otimes \mathcal{A}) \to (M, \mathcal{B}(M)),$$
$$(x, \omega) \mapsto \varphi_{s,t}(x, \omega)$$

is measurable and if it satisfies the following properties.

(a) (Cocycle property) For all $s < u < t$ and $x \in M$, $\mathsf{P}$-almost surely, $\varphi_{s,t}(x) = \varphi_{u,t} \circ \varphi_{s,u}(x)$.

(b) (Stationarity) For all $s \le t$, $\varphi_{s+h,t+h} = \varphi_{s,t} \circ T_h$.

(c) The flow has independent increments, i.e. for all $t_1 < t_2 < \cdots < t_n$, the family $\{\varphi_{t_i,t_{i+1}},\ 1 \le i \le n-1\}$ is independent.

(d) For every $f \in C(M)$, $\lim_{(u,v)\to(s,t)} \sup_{x\in M} \mathsf{E}[(f\circ\varphi_{s,t}(x) - f\circ\varphi_{u,v}(x))^2] = 0$.

(e) For all $f \in C(M)$ and $s \le t$, $\lim_{d(x,y)\to 0} \mathsf{E}[(f \circ \varphi_{s,t}(x) - f \circ \varphi_{s,t}(y))^2] = 0$.

Let $\varphi = (\varphi_{s,t},\ s \le t)$ be a stochastic flow of mappings. For all $n \ge 1$, $f \in C(M^n)$ and $x \in M^n$ set

$$\mathsf{P}_t^{(n)} f(x) = \mathsf{E}[f \circ \varphi_{0,t}^{\otimes n}(x)]. \tag{1.7}$$

Then $(\mathsf{P}_t^{(n)},\ n \ge 1)$ is a compatible family of Feller semigroups on $M$ satisfying (1.6). *The law of $\varphi$ is uniquely determined by* $(\mathsf{P}_t^{(n)},\ n \ge 1)$.

**Theorem 1.1.5.** 1) *Let* $(\mathsf{P}_t^{(n)},\ n \ge 1)$ *be a compatible family of Feller semigroups on $M$ satisfying*

$$\mathsf{P}_t^{(2)} f^{\otimes 2}(x, x) = \mathsf{P}_t f^2(x) \tag{1.8}$$

*for all $f \in C(M)$, $x \in M$ and $t \ge 0$. Then there exists a unique Feller convolution semigroup $(\mathsf{Q}_t)_{t\ge 0}$ on $(F, \mathcal{F})$ such that for all $n \ge 1$, $t \ge 0$, $f \in C(M^n)$ and $x \in M^n$,*

$$\mathsf{P}_t^{(n)} f(x) = \int f \circ \varphi^{\otimes n}(x)\ \mathsf{Q}_t(d\varphi). \tag{1.9}$$

2) *For every Feller convolution semigroup $\mathsf{Q} = (\mathsf{Q}_t)_{t\ge 0}$ on $(F, \mathcal{F})$ there exists a stochastic flow of mappings associated with $\mathsf{Q}$ (or equivalently with* $(\mathsf{P}_t^{(n)},\ n \ge 1)$*).*

Let $V, V_1, \ldots, V_k$ be bounded Lipschitz vector fields on a smooth locally compact manifold $M$. We also assume that $V_1, \ldots, V_k$ are $C^1$. Let $W^1, \ldots, W^k$ be $k$ independent real white noises. We consider the following SDE on $M$:

$$dX_t = \sum_{i=1}^{k} V_i(X_t) \circ dW_t^i + V(X_t)\, dt, \quad t \in \mathbb{R}. \tag{1.10}$$

From the usual theory of strong solutions of SDEs (see for example [12]) it is possible to construct a stochastic flow of diffeomorphisms $(\varphi_{s,t},\ s \le t)$ such that for every $x \in M$, $\varphi_{s,t}(x)$ is a strong solution of the SDE (1.10) with $\varphi_{s,s}(x) = x$.

Using this stochastic flow, it is possible to construct a compatible family of Markovian semigroups $(\mathsf{P}_t^{(n)},\ n \ge 1)$ with

$$\mathsf{P}_t^{(n)} h(x_1, \dots, x_n) = \mathsf{E}[h(\varphi_{0,t}(x_1), \dots, \varphi_{0,t}(x_n))] \tag{1.11}$$

for $h \in C(M^n)$ and $x_1, \dots, x_n$ in $M$. It is easy to check that these semigroups are Feller and that the canonical stochastic flow of maps associated with this family of semigroups is equal in law to $(\varphi_{s,t},\ s \le t)$.

**1.2. Flows of transition kernels.** We denote by $\mathcal{P}(M)$ the space of probability measures on $M$, equipped with the weak convergence topology. $\mathcal{P}(M)$ is a compact metric space. Let us recall that a kernel $K$ on $M$ is a measurable mapping from $M$ into $\mathcal{P}(M)$, $M$ and $\mathcal{P}(M)$ being equipped with their Borel $\sigma$-fields. For all $f \in C(M)$ and $x \in M$, $Kf(x)$ denotes $\int f(y)\, K(x, dy)$. For every $\mu \in \mathcal{P}(M)$, $\mu K$ denotes the probability measure defined by $\int f(y)\, \mu K(dy) = \int Kf(x)\, \mu(dx)$. We denote by $E$ the space of all kernels on $M$, and we equip $E$ with the $\sigma$-field $\mathcal{E}$ generated by the mappings $K \mapsto \mu K$ for every $\mu \in \mathcal{P}(M)$. Convolution semigroups on the space of kernels can be defined in a similar way as on the space of measurable maps (cf. [16]).

**Definition 1.2.1.** Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and $(T_h)_{h\in\mathbb{R}}$ a one parameter group of $\mathsf{P}$-preserving $L^2$-continuous transformations of $\Omega$. Then a family of $(E, \mathcal{E})$-valued random variables $(K_{s,t},\ s \le t)$ is called a (*measurable*) *stochastic flow of kernels* if for all $s \le t$,

$$(x, \omega) \mapsto K_{s,t}(x, \omega) \tag{1.12}$$

is a measurable mapping from $(M \times \Omega, \mathcal{B}(M) \otimes \mathcal{A})$ onto $(\mathcal{P}(M), \mathcal{B}(\mathcal{P}(M)))$ and if it satisfies the following properties.

(a) (Cocycle property) For all $s < u < t$ and $x \in M$, $\mathsf{P}$-almost surely, for every $f \in C(M)$, $K_{s,t} f(x) = K_{s,u}(K_{u,t} f)(x)$.

(b) (Stationarity) For all $s \le t$, $K_{s+h,t+h} = K_{s,t} \circ T_h$.

(c) The flow has independent increments, i.e. for all $t_1 < t_2 < \cdots < t_n$, the family $\{K_{t_i,t_{i+1}},\ 1 \le i \le n-1\}$ is independent.

(d) For every $f \in C(M)$,

$$\lim_{(u,v)\to(s,t)} \sup_{x\in M} \mathsf{E}[(K_{s,t} f(x) - K_{u,v} f(x))^2] = 0. \tag{1.13}$$

(e) For all $f \in C(M)$ and $s < t$,

$$\lim_{d(x,y)\to 0} \mathsf{E}[(K_{s,t} f(x) - K_{s,t} f(y))^2] = 0. \tag{1.14}$$

Let $(K_{s,t},\ s \le t)$ be a stochastic flow of kernels. For all $n \ge 1$, $f \in C(M^n)$ and $x \in M^n$ set

$$\mathsf{P}_t^{(n)} f(x) = \mathsf{E}[K^{\otimes n} f(x)]. \tag{1.15}$$

Then $(\mathsf{P}_t^{(n)},\ n \ge 1)$ is a compatible family of Feller semigroups on $M$.

**Theorem 1.2.2.** 1) *For every compatible family* $(\mathsf{P}_t^{(n)},\ n \ge 1)$ *of Feller semigroups on* $M$ *there exists a unique Feller convolution semigroup* $(\nu_t)_{t\ge 0}$ *on* $(E, \mathcal{E})$ *such that for all* $n \ge 1$, $t \ge 0$, $f \in C(M^n)$ *and* $x \in M^n$,

$$\mathsf{P}_t^{(n)} f(x) = \int K^{\otimes n} f(x)\ \nu_t(dK). \tag{1.16}$$

2) *For every Feller convolution semigroup* $\nu = (\nu_t)_{t\ge 0}$ *on* $(E, \mathcal{E})$ *there exists a stochastic flow of kernels associated with* $\nu$ (*or equivalently with* $(\mathsf{P}_t^{(n)},\ n \ge 1)$).

**Remark 1.2.3.** If (1.6) is satisfied the stochastic flow of kernels $K$ is induced by a stochastic flow of mappings $\varphi$.

## 2. Noise and stochastic flows

The definition of a noise we give here is very close to the one given by Tsirelson in [25].

**Definition 2.1.1.** A *noise* consists of a separable probability space $(\Omega, \mathcal{A}, \mathsf{P})$, a one parameter group $(T_h)_{h\in\mathbb{R}}$ of $\mathsf{P}$-preserving $L^2$-continuous transformations of $\Omega$ and a family $\{\mathcal{F}_{s,t},\ -\infty \le s \le t \le \infty\}$ of sub-$\sigma$-fields of $\mathcal{A}$ such that

(a) $T_h$ maps $\mathcal{F}_{s,t}$ onto $\mathcal{F}_{s+h,t+h}$ for all $h \in \mathbb{R}$ and $s \le t$,

(b) $\mathcal{F}_{s,t}$ and $\mathcal{F}_{t,u}$ are independent for all $s \le t \le u$,

(c) $\mathcal{F}_{s,t} \vee \mathcal{F}_{t,u} = \mathcal{F}_{s,u}$ for all $s \le t \le u$.

A classical white noise or a stationary Poisson measure clearly define a noise in this sense.

A square integrable random variable with zero mean is said to belong to the first chaos if the sum of its conditional expectations with respect to the fields associated with disjoint intervals is the conditional expectation with respect to the field associated with the union of these intervals. The noise is called black when the first chaos reduces to zero. Clearly, the white noise or the Poisson noise are not black.

Let $(\Omega, \mathcal{A}, \mathsf{P}_\nu)$ denote the probability space of a stochastic flow of kernels $K = (K_{s,t},\ s \le t)$ associated with a Feller convolution semigroup $\nu$.

For all $-\infty \le s \le t \le \infty$ let $\mathcal{F}_{s,t}$ be the sub-$\sigma$-field of $\mathcal{A}$ generated by the random variables $K_{u,v}$ for all $s \le u \le v \le t$. Then the cocycle property of $K$ implies that $N_\nu := (\Omega, \mathcal{A}, (\mathcal{F}_{s,t})_{s\le t}, \mathsf{P}_\nu, (T_h)_{h\in\mathbb{R}})$ is a noise ($T_h$ is $L^2$-continuous by the Feller property). We call it the noise generated by the flow $K$.

## 3. Stochastic coalescing flows

Starting from a compatible family of Feller semigroups, under the hypothesis that the two-point motion hits the diagonal almost surely, we construct another compatible family of Feller semigroups to which is associated a stochastic coalescing flow. It appears that the stochastic flow of kernels associated with the first family of semigroups can be recovered by filtering the stochastic coalescing flow with respect to a sub-noise of an extension of its noise.

Finally, we give the example of Arratia's flow ([2]), which describes a space-time continuum of independent Brownian motions sticking together when they meet. The construction of a stochastic coalescing flow solution of SDE's will be presented in the next sections (see also [11], [5]).

### 3.1. Definition and construction

**Definition 3.1.1.** A stochastic flow of mappings on $M$, $(\varphi_{s,t},\ s \leq t)$, is called a *stochastic coalescing flow* if for all $(x, y) \in M^2$, $T_{x,y} = \inf\{t \geq 0,\ \varphi_{0,t}(x) = \varphi_{0,t}(y)\}$ is finite and for every $t \geq T_{x,y}$, $\varphi_{0,t}(x) = \varphi_{0,t}(y)$ almost surely.

This definition depends only on the two-point motion.

Let $(\mathsf{P}_t^{(n)},\ n \geq 1)$ be a compatible family of Feller semigroups on a compact separable metric space $M$, $\nu = (\nu_t)_{t\in\mathbb{R}}$ the associated Feller convolution semigroup on $(E, \mathcal{E})$ and $K_t$ the associated flow of kernels. Let $\Delta_n = \{x \in M^n,$ there exists $i \neq j$ such that $x_i = x_j\}$ and $T_{\Delta_n} = \inf\{t \geq 0,\ X_t^{(n)} \in \Delta_n\}$, where $X_t^{(n)}$ denotes the $n$-point motion, i.e. the Markov process on $M^n$ associated with the semigroup $\mathsf{P}_t^{(n)}$. Denoting by $\mathsf{P}_{(x,y)}^{(2)}$ the law of the Markov process associated with $\mathsf{P}_t^{(2)}$ starting from $(x, y)$ and $\Delta_2$ by $\Delta$, assume that for all $t > 0$ and $x \in M$,

$$\lim_{y\to x} \mathsf{P}_{(x,y)}^{(2)}[\{T_\Delta > t\}] = 0,$$

and that for all $x$ and $y$ in $M$, $\mathsf{P}_{(x,y)}^{(2)}[T_\Delta < \infty] = 1$.

**Theorem 3.1.2.** *There exists a unique compatible family* $(\mathsf{P}_t^{(n),c},\ n \geq 1)$ *of Feller semigroups on* $M$ *such that if* $X^{(n),c}$ *is the associated* $n$*-point motion and* $T_{\Delta_n}^c = \inf\{t \geq 0,\ X_t^{(n),c} \in \Delta_n\}$, *then*

- $(X_t^{(n),c},\ t \leq T_{\Delta_n}^c)$ *is equal in law to* $(X_t^{(n)},\ t \leq T_{\Delta_n})$,
- *for* $t \geq T_{\Delta_n}^c$, $X_t^{(n),c} \in \Delta_n$.

*Moreover,* $(\mathsf{P}_t^{(n),c},\ n \geq 1)$ *satisfies* (1.6) *and is associated with a coalescing flow* $\varphi_{s,t}^c$.

We denote by $\nu^c$ the associated Feller convolution semigroup. An important result is the following:

**Theorem 3.1.3.** *There is a joint realisation of $K$ and $\varphi^c$ such that $K_{s,t}g(y) = \mathsf{E}[\varphi^c_{s,t}(y)|\sigma(K)]$*

We say that the convolution semigroup $\nu^c$ weakly dominates $\nu$.

**3.2. Arratia's coalescing flow of independent Brownian motions.** The first example of coalescing flows was given by Arratia [2]. On $\mathbb{R}$, or on the unit circle, let $\mathsf{P}_t$ be the semigroup of a Brownian motion. With this semigroup we define the compatible family $(\mathsf{P}_t^{\otimes n},\ n \geq 1)$ of Feller semigroups. Note that the $n$-point motion of this family of semigroups is given by $n$ independent Brownian motions. Let us also remark that the canonical stochastic flow of kernels associated with this family of semigroups is not random and is given by $(\mathsf{P}_{t-s},\ s \leq t)$.

Let $(\mathsf{P}_t^{(n)},\ n \geq 1)$ be the compatible family of Markovian coalescent semigroups associated with $(\mathsf{P}_t^{\otimes n},\ n \geq 1)$. Note that the $n$-point motion of this family of semigroups is given by $n$ independent Brownian motions which stick together when they meet.

**Theorem 3.2.1.** *The family $(\mathsf{P}_t^{(n)},\ n \geq 1)$ is constituted of Feller semigroups and is associated with a coalescing flow. The noise defined by this flow is black*

Blackness of the noise was first proved in [25] and then in a different way in [17]. It may seem a paradox but note that the increments of a one point motion between two times depend on the position of that point at the first time and not only on the increment of the flow of maps. In the latter paper a related family of flows of kernels, called sticky flows, was also constructed. The associated $n$ point motions are given by Brownian paths which are independent except they stick together (during a Cantor type set of times of positive Lebesgue measure) when they meet. These flows interpolate between the heat flow and Arratia's flow. They also define a black noise. Any flow of kernels induces naturally a Markov process on measures which has often an invariant probability distribution. For sticky flows it is explicitly given in terms of the Poisson Dirichlet distribution. Finally, let us mention that a discrete model converging to these flows was presented in [13].

## 4. Tanaka's equation

Tanaka's stochastic differential equation (SDE) is one of the simplest examples of an SDE that does not have a strong solution in the usual sense. The objective is to apply to this example the theory of flows of transition kernels and to classify all the solutions of Tanaka's SDE, extended to transition kernels. It is shown that they can be characterized by a probability measure on $[0, 1]$. The domination and the weak domination relations (defined in [15]) between different solutions are then fully understood in terms of barycenter and balayage of the associated measures.

On a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ let $W = (W_{s,t},\ s \le t)$ be a real white noise and $K = (K_{s,t},\ s \le t)$ (resp. $\varphi = (\varphi_{s,t},\ s \le t)$) be a stochastic flow of kernels (resp. flow of measurable maps) on the real line. Recall that for all $s \le t$, $K_{s,t} : \mathbb{R} \to \mathcal{P}(\mathbb{R})$ is measurable, with $\mathcal{P}(\mathbb{R})$ denoting the set of probability measures on $\mathbb{R}$ equipped with the topology of weak convergence. We say that $(K, W)$ solves Tanaka's SDE if for all $s \le t$, $f \in C_K^2(\mathbb{R})$ and $x \in \mathbb{R}$,

$$K_{s,t} f(x) = f(x) + \int_s^t K_{s,u}(f' \mathrm{sgn})(x) W(du) + \frac{1}{2} \int_s^t K_{s,u}(f'') du, \tag{4.1}$$

with $\mathrm{sgn}(x) = 1_{x \ge 0} - 1_{x<0}$. Note that (4.1) is a generalization of the SDE

$$dX_t = \mathrm{sgn}(X_t) dW_t,$$

where $W_t = W_{0,t} 1_{t \ge 0} - W_{t,0} 1_{t<0}$.

It can be shown that this implies that $\sigma(W) \subset \sigma(K)$. Let $N^K$ be the noise of $K$. The noise $N^W$ of $W$ is a subnoise of $N^K$. So we can simply say that $K$ solves Tanaka's SDE (since $W$ is a function of $K$). We say that a flow of maps $\varphi$ solves Tanaka's SDE if $\delta_\varphi$ solves Tanaka's SDE. The law of a solution $K$ is given by a Feller convolution semigroup $\nu = (\nu_t,\ t \ge 0)$, where $\nu_t$ is the law of $K_{0,t}$.

Two particular solutions of Tanaka's SDE are given in [15]: the coalescing solution $\varphi^c$ and the Wiener solution $K^W$. The solution $K^W$ is the only solution of Tanaka's SDE such that $N^K = N^W$, and $\varphi^c$ is the only flow of maps solution of Tanaka's SDE. The Wiener solution can be obtained by filtering the coalescing solution: $K^W = \mathsf{E}[\delta_\varphi | W]$. An explicit expression of $K^W$ can be given. For $x \in \mathbb{R}$ set $\tau_x = \inf\{t > 0,\ W_{0,t} = -|x|\}$. Let $W^+ = (W_t^+,\ t \ge 0)$ be defined by

$$W_t^+ = W_{0,t} - \inf_{s \le t} W_{0,s}.$$

It is well known that the law of $(W_t^+)_{t \ge 0}$ and the law of $(|W_t|)_{t \ge 0}$ coincide. Note that $W_{0,\cdot}$ can be recovered out of $W^+$ by Doob–Meyer decomposition. Then for $t \ge 0$,

$$K_{0,t}^W(x) = \delta_{x + \mathrm{sgn}(x) W_{0,t}} 1_{\{t \le \tau_x\}} + \frac{1}{2}(\delta_{W_t^+} + \delta_{-W_t^+}) 1_{\{t > \tau_x\}}. \tag{4.2}$$

Let $\theta_h^W$ be the shift operator such that $W_{s,t} \circ \theta_h^W = W_{s+h,t+h}$. Then for all $s < t$, $K_{s,t}^W = K_{0,t-s}^W \circ \theta_s^W$. The coalescing solution $\varphi^c$ can be defined by the consistent family of its $n$-point motions obtained by transforming the $n$-point motion associated with $K^W$ into a coalescing motion. A more explicit definition can be given in this special case, as is shown in [18], where we also prove the following result:

**Theorem 4.1.1.** a) *Each solution $K$ of Tanaka's SDE verifies $K^W = \mathsf{E}[K|W]$ (this means in particular that the support of $K$ has at most two points). It defines a probability measure $m$ on $[0, 1]$ with mean $1/2$, which is the law of $\int_0^\infty K_{0,t}(0, dy)$ for all $t > 0$.*

b) *The mapping defined in* a) *is a bijection between solutions of* (4.1) *and probability measures on* $[0, 1]$ *with mean* $1/2$. *The Feller convolution semigroup associated with a measure* $m$ *is denoted* $\{\nu_t^m,\ t \geq 0\}$ *or* $\nu^m$.

c) $K^W$ *is associated with* $\delta_{1/2}$ *and* $\varphi^c$ *with* $\frac{1}{2}(\delta_0 + \delta_1)$.

Let us now describe the domination relations.

**Definition 4.1.2.** Let $m_1$ and $m_2$ be probability measures on $[0, 1]$.

a) $m_1$ is swept by $m_2$ if and only if for all positive convex function $f$, $\int f dm_2 \leq \int f dm_1$.

b) $m_2$ is a barycenter of $m_1$ if and only if there exists a measurable map $\psi : [0, 1] \to [0, 1]$ such that $\psi^* m_1 = m_2$ and $\psi^*(I \cdot m_1) = I \cdot m_2$ (where $I$ denotes the identity function).

It can easily be seen that a) and b) define partial order relations. The order defined in a) is the balayage order. The fact that $m_2$ is a barycenter of $m_1$ is equivalent to saying that if $U_1$ is a random variable of law $m_1$, then there exists a $\sigma(U_1)$-measurable random variable $U_2$ of law $m_2$ such that $\mathsf{E}[U_1|U_2] = U_2$.

In [16], a domination and a weak domination relation between (laws of) stochastic flow of kernels is defined: Let $\nu^1$ and $\nu^2$ be two Feller convolution semigroups. We recall that Definition 3.3 in [16] essentially says that $\nu^1$ dominates $\nu^2$ if and only if there is a joint realisation $(K^1, K^2)$ such that $K^1$ (resp. $K^2$) is a stochastic flow of kernels associated to $\nu^1$ (resp. to $\nu^2$) satisfying $\mathsf{E}[K^1|K^2] = K^2$ and $\sigma(K^2) \subset \sigma(K^1)$. One says that $\nu^1$ weakly dominates $\nu^2$ when only the conditional expectation assumption is verified ($\sigma(K^2)$ needs not be a sub-$\sigma$-field of $\sigma(K^1)$). A full understanding of the solutions of a general SDE should involve a classification of the solutions according to these domination relations. As we will see in the following section, this is not achieved yet even in relatively simple cases.

**Theorem 4.1.3.** *Let* $m_1$ *and* $m_2$ *be two probability measures on* $[0, 1]$ *with mean* $1/2$.

a) $\nu^{m_1}$ *dominates* $\nu^{m_2}$ *if and only if* $m_2$ *is a barycenter of* $m_1$.

b) $\nu^{m_1}$ *weakly dominates* $\nu^{m_2}$ *if and only if* $m_1$ *is swept by* $m_2$.

## 5. Stochastic flows of kernels and SDEs: an example on the circle

**Notation.** In all the following we will denote by $\mathbb{S}$ the unit circle $\mathbb{R}/2\pi\mathbb{Z}$, by $m$ the Lebesgue measure on $\mathbb{S}$ and by $\mathcal{P}(\mathbb{S})$ the set of Borel probability measures on $\mathbb{S}$.

Let $(\mathcal{W}, \mathcal{F}^W, \mathsf{P}_W)$ be the canonical probability space of a sequence of independent Wiener processes $(W_t^k,\ k \geq 0,\ t \geq 0)$. For all $s < t$ let $\mathcal{F}_{s,t}^W$ denote the $\sigma$-field generated by the random variables $W_v^k - W_u^k$, $s \leq u < v \leq t$ and $k \geq 0$. Being

given $(a_k)_{k\geq 0}$ a sequence of nonnegative numbers such that $\sum_{k\geq 0} a_k^2 < \infty$, we set $C(z) = \sum_{k\geq 0} a_k^2 \cos(kz)$. Note that all real positive definite functions on $\mathbb{S}$ can be written in this form and that $C(0) = \sum_{k\geq 0} a_k^2$.

**5.1. Flows of diffeomorphisms.** Assume that $\sum_{k\geq 1} k^2 a_k^2 < \infty$. Then by a stochastic version of Gronwall's lemma it can be shown that for each $x_0 \in \mathbb{S}$ the stochastic differential equation (SDE)

$$x_t = x_0 + a_0 W_t^0 + \sum_{k\geq 1} a_k \left( \int_0^t \sin(kx_s) dW_s^{2k-1} + \int_0^t \cos(kx_s) dW_s^{2k} \right) \tag{5.1}$$

has a unique strong solution. These solutions can be considered jointly to form a stochastic flow of diffeomorphisms $(\varphi_{s,t})_{s<t}$. Set $\varphi_t = \varphi_{0,t}$.

Note that the one point motion $x_t := \varphi_t(x)$ is a Brownian motion on $\mathbb{S}$ starting at $x$. Denote the associated heat semigroup $\mathsf{P}_t$. For $h \in C^2(\mathbb{S}^2)$ set $A^{(2)}h(x, y) = \frac{C(0)}{2}(\partial_{xx}^2 h(x, y) + \partial_{yy}^2 h(x, y)) + C(x-y)\partial_{xy}^2 h(x, y)$. The two point motion $(x_t, y_t) := (\varphi_t(x), \varphi_t(y))$ is a diffusion on $\mathbb{S}^2$ satisfying

- for all $h \in C^2(\mathbb{S}^2)$, $h(x_t, y_t) - \int_0^t A^{(2)}h(x_s, y_s)ds$ is an $L^2$-martingale.

This in particular implies the following:

- For all $x, y \in \mathbb{S}$, $f \in C(\mathbb{S})$, if $Z_t = \varphi_t(x) - \varphi_t(y)$, $Z_t$ is a diffusion on $\mathbb{S}$ and $f(Z_t) - \int_0^t (C(0) - C(Z_s)) f''(Z_s) ds$ is an $L^2$-martingale.

For all $x, y \in \mathbb{S}$, $\lim_{t\to\infty}(\varphi_t(x) - \varphi_t(y)) = 0$. Using the isotropy one can compute the Lyapounov exponent of the flow: for all $x \in \mathbb{S}$,

$$\lim_{t\to\infty} \frac{1}{t} \log |\varphi_t'(x)| = -\frac{1}{2} \sum_{k\geq 1} k^2 a_k^2.$$

The Lyapounov exponent of the flow being negative, the flow is stable. In the particular case $a_k^2 = k^{-(1+\alpha)}$, for $\alpha > 2$ the condition $\sum_{k\geq 1} k^2 a_k^2 < \infty$ is satisfied. When $\alpha$ is close to the boundary case, $\alpha = 2$, the Lyapounov exponent gets close to $-\infty$. In the following we will define stochastic flows corresponding to the case $\alpha \leq 2$ which are not flows of diffeomorphisms.

**5.2. Wiener chaos expansion: Lipschitz case.** Suppose that $\sum_k k^2 a_k^2 < \infty$ and let $\varphi_{s,t}$ be the flow defined in the previous section. For any function $f \in C(\mathbb{S})$, $x \in \mathbb{S}$ and $s \leq t$, $f \circ \varphi_{s,t}(x)$ belongs to the Wiener space $L^2(\mathsf{P}_W)$. Following the original idea of [26], its Wiener chaos expansion can be explicitly computed as follows.

**Proposition 5.2.1.** *For all $s \le t$ and $f \in C(\mathbb{S})$,*

$$f \circ \varphi_{s,t}(x) = \mathsf{P}_{t-s} f(x) + \sum_{n \ge 1} J^n_{s,t} f(x) \quad \text{in } L^2(\mathsf{P}_W), \tag{5.2}$$

*where $J^n$ is defined recursively as follows (denoting $c_k$ the function $x \mapsto \cos(kx)$ and $s_k$ the function $x \mapsto \sin(kx)$):*

$$\begin{aligned} J^{n+1}_{s,t} f(x) = a_0 \int_s^t J^n_{s,u}((\mathsf{P}_{t-u} f)')(x) dW^0_u \\ + \sum_{k \ge 1} a_k \Big( \int_s^t J^n_{s,u}(s_k(\mathsf{P}_{t-u} f)')(x) dW^{2k-1}_u \\ + \int_s^t J^n_{s,u}(c_k(\mathsf{P}_{t-u} f)')(x) dW^{2k}_u \Big) \end{aligned}$$

*for $n \ge 0$ with $J^0_{s,t} = \mathsf{P}_{t-s}$.*

**Remark 5.2.2.** The chaos expansion (5.2) can be extended to all $f \in L^2(m)$, the two terms being equal in $L^2(m \otimes \mathsf{P}_W)$.

**5.3. Non-Lipschitz case.** From now on we assume $\sum_{k \in \mathbb{N}} k^2 a_k^2 = \infty$. In this case, using Gronwall's inequality, the existence of a strong solution to the SDE (5.1) cannot be proven. But the series giving the Wiener chaos expansion of $f \circ \varphi_{s,t}$ in the Lipschitz case for $f \in L^2(m)$ also converges in $L^2(m \otimes \mathsf{P}_W)$ in the non-Lipschitz case.

We can construct a family $S^n_{s,t}$ of random operators acting of $L^2(m)$ recursively: let $S^0_{s,t} = \mathsf{P}_{t-s}$ and for $f \in L^2(m)$ and $n \ge 0$ set

$$\begin{aligned} S^{n+1}_{s,t} f = \mathsf{P}_{t-s} f + a_0 \int_0^t S^n_{s,u}((\mathsf{P}_{t-u} f)') dW^0_u \\ + \sum_{k \ge 1} a_k \left( \int_s^t S^n_{s,u}(s_k(\mathsf{P}_{t-u} f)') dW^{2k-1}_u + \int_s^t S^n_{s,u}(c_k(\mathsf{P}_{t-u} f)') dW^{2k}_u \right). \end{aligned}$$

It can be seen that for all $n$, $\mathsf{E}[(S^n_{s,t} f)^2] \le \mathsf{P}_{t-s} f^2$ and

$$S^n_{s,t} f = \sum_{k=0}^n J^k_{s,t} f, \tag{5.3}$$

where $J^k_{s,t} f$ belongs to the $k$-th Wiener chaos. Thus all these terms are orthogonal and $S^n_{s,t} f$ converges in $L^2(m \otimes \mathsf{P}_W)$ towards a limit we denote by $S_{s,t} f$. The family $S = (S_{s,t})$ of random operators acting on $L^2(m)$ satisfies the following.

(i) Cocycle property: $S_{s,u} = S_{s,t} S_{t,u}$ for all $s < t < u$.

(ii) Stationary increments: for all $s \le t$, $S_{s,t}$ and $S_{0,t-s}$ have the same law.

(iii) Independent increments: for $t_0 \le \cdots \le t_n$, $S_{t_0,t_1}, \ldots, S_{t_{n-1},t_n}$ are independent.

(iv) Solution of the SDE

$$\begin{aligned} S_{s,t}f = f &+ a_0 \int_s^t S_{s,u}(f')dW_u^0 \\ &+ \sum_{k\ge 1} a_k \left( \int_s^t S_{s,u}(s_k f')dW_u^{2k-1} + \int_s^t S_{s,u}(c_k f')dW_u^{2k} \right) \\ &+ \frac{C(0)}{2} \int_s^t S_{s,u} f'' du, \end{aligned} \tag{5.4}$$

for all $f \in H^2(\mathbb{S})$ and all $s < t$.

Moreover, $S$ is the unique family of random operators acting on $L^2(m)$ verifying $\mathsf{E}[(S_{s,t}f)^2] \le \mathsf{P}_{t-s}f^2$, satisfying (i), (ii), (iii), (iv) and such that $S_{s,t}$ is $\mathcal{F}_{s,t}^W$-measurable.

Obviously $S_{s,t}1 = 1$, and it can be proved that $S_{s,t}$ is nonnegative as follows. Consider an independent stationary Brownian motion $B_t$ with diffusion coefficient $C(0)$ on $\mathbb{S}$. Set, for $k \ge 1$,

$$\begin{aligned} \widetilde{W}_t^{2k-1} = W_t^{2k-1} &+ a_k \int_0^t s_k(B_s)dB_s - a_k a_0 \int_0^t s_k(B_s)dW_s^0 \\ &- a_k \sum_{l\ge 1} a_l \left( \int_0^t s_k s_l(B_s)dW_s^{2l-1} + \int_0^t s_k c_l(B_s)dW_s^{2l} \right) \end{aligned}$$

and, for $k \ge 0$,

$$\begin{aligned} \widetilde{W}_t^{2k} = W_t^{2k} &+ a_k \int_0^t c_k(B_s)dB_s - a_k a_0 \int_0^t c_k(B_s)dW_s^0 \\ &- a_k \sum_{l\ge 1} a_l \left( \int_0^t c_k s_l(B_s)dW_s^{2l-1} + \int_0^t c_k c_l(B_s)dW_s^{2l} \right). \end{aligned}$$

These formulas are obtained by conditioning the "velocity differential" at time $t$ and site $B_t$ to be $dB_t$. Then $\widetilde{W}$ forms a family of independent Wiener processes. Set

$$\tilde{S}_{s,t}f(x) = \mathsf{E}[f(B_t)|\widetilde{W}, B_s = x].$$

It is clear that $\tilde{S}$ is nonnegative and that $\tilde{S}$ verifies the properties listed above ((i), (ii), (iii) and (iv)) with respect to $\widetilde{W}$. This implies $\tilde{S} = S$ and proves that $S$ is nonnegative.

Two cases may occur:

(a) $S_{s,t}f^2 = (S_{s,t}f)^2$ for all $f \in L^\infty(m)$.

(b) $S_{s,t}f^2 > (S_{s,t}f)^2$ for some $f \in L^\infty(m)$, and in fact for all non constant $f \in L^\infty(m)$.

**5.4. *n*-point motions.** Let $\mathsf{P}_t^{(n)}$ be the family of random operators acting on $L^\infty(m^{\otimes n})$ defined by

$$\mathsf{P}_t^{(n)} f_1 \otimes \cdots \otimes f_n = \mathsf{E}[S_{0,t} f_1 \otimes \cdots \otimes S_{0,t} f_n].$$

Properties (i), (ii) and (iii) imply that $\mathsf{P}_t^{(n)}$ is a Markovian semigroup. As in the case of $\mathbb{R}^d$ or $\mathbb{S}^d$ studied in [16], one can show that the isotropy implies that $\mathsf{P}_t^{(n)}$ is a Feller semigroup acting on $C(\mathbb{S}^n)$.

The *n-point motion* of $(S_{s,t})$ is the diffusion on $\mathbb{S}^n$ associated with $\mathsf{P}_t^{(n)}$. The generator $A^{(n)}$ of this diffusion is given by

$$A^{(n)} = \frac{1}{2} \sum_{1 \le i,j \le n} C(x_i - x_j) \partial_{x_i} \partial_{x_j}. \tag{5.5}$$

The case (a) appears when the diagonal is absorbing for the two-point motion. If this is not the case we are in case (b).

**5.5. Diffusive or coalescing?** In case (a) it can be shown (using the Feller property) that there exists a flow of random mappings $\varphi = (\varphi_{s,t})$ such that for all $s \le t$ and all $f \in L^2(m)$, we have $S_{s,t} f = f \circ \varphi_{s,t}$ in $L^2(m \otimes \mathsf{P}_W)$. Furthermore, $\varphi_{s,t} \colon (\mathbb{S} \times \mathcal{W}, \mathcal{B}(\mathbb{S}) \otimes \mathcal{F}^W) \to (\mathbb{S}, \mathcal{B}(\mathbb{S}))$ is measurable and solves the SDE (5.1).

In case (b) it can be shown that there exists a flow of random kernels $K = (K^W_{s,t})$ such that $S_{s,t} f = K^W_{s,t} f$ in $L^2(m \otimes \mathsf{P}_W)$ for all $s \le t$ and all $f \in L^2(m)$. The stochastic flow of kernels will be called *diffusive* when the kernels are not induced by maps, which clearly happens in case (b). This flow solves the SDE in the sense that for all $f \in C^2(\mathbb{S})$, $s \le t$ and $x \in \mathbb{S}$,

$$\begin{aligned} K^W_{s,t} f = f + \sum_{k \ge 1} a_k \left( \int_s^t K^W_{s,u}(s_k f') dW_u^{2k-1} + \int_s^t K^W_{s,u}(c_k f') dW_u^{2k} \right) \\ + a_0 \int_s^t K^W_{s,u}(f') dW_u^0 \quad + \quad \frac{C(0)}{2} \int_s^t K^W_{s,u} f'' du. \end{aligned} \tag{5.6}$$

In the following the flow $\varphi$ (in case (a)) or the flow $K^W$ (in case (b)) will be called the *Wiener solution* of the SDE (5.1). Since $(S_{s,t})$ is the unique solution of (5.4) which is $\mathcal{F}^W_{s,t}$-measurable, the Wiener solution $\varphi$ (or $K^W$) is the unique solution of SDE (5.1) (or of (5.6)) which is $\mathcal{F}^W_{s,t}$-measurable.

A diffusive flow is called *diffusive with hitting* if the two-point motion hits the diagonal $\Delta = \{(x, x),\ x \in \mathbb{S}\}$.

The diffusion $z_t \in [0, 2\pi)$ such that $z_t = X_t - Y_t$ modulo $2\pi$, where $(X_t, Y_t)$ is the two point motion, has a natural scale. The speed measure $m$ of this diffusion is given by $m(dz) = (C(0) - C(z))^{-1} dz$. Let $\kappa$ be defined by $\kappa(z) = \int_\pi^z \frac{z-x}{C(0)-C(x)} dx$. Note that $\kappa(0+) = \infty$ implies that $m((0, 2\pi)) = \infty$.

**Theorem 5.5.1.** 1) *If $\kappa(0^+) = \infty$ then the Wiener solution is a stochastic flow of maps, which is not a coalescing flow.*

2) *If $m((0, 2\pi)) = \infty$ and $\kappa(0^+) < \infty$ then the Wiener solution is a coalescing flow.*

3) *If $m((0, 2\pi)) < \infty$ then the Wiener solution is a diffusive flow with hitting.*

**Corollary 5.5.2.** *Let $a_k^2 = k^{-(1+\alpha)}$ with $\alpha > 0$.*

1) *If $\alpha > 2$, then the Wiener solution is a stochastic flow of $C^1$-diffeomorphisms.*

2) *If $\alpha = 2$ then the Wiener solution is a stochastic flow of maps, which is not a coalescing flow.*

3) *If $\alpha \in [1, 2)$ then the Wiener solution is a coalescing flow.*

4) *If $\alpha \in (0, 1)$ then the Wiener solution is a diffusive flow with hitting.*

**Remark 5.5.3.** The case $\alpha = 2$ has been studied in [1], [8], [22]. It is shown in particular that the maps of the flow are homeomorphisms.

**5.6. Extension of the noise and weak solution.** Let $(\Omega, \mathcal{A}, \mathsf{P})$ be an extension of the probability space $(\mathcal{W}, \mathcal{F}^W, \mathsf{P}_W)$. We say that a measurable flow of maps $\varphi = (\varphi_{s,t})$ is a *weak solution* of (5.1) if it satisfies (5.1) without being $\mathcal{F}_{s,t}^W$-measurable. Similarly, a measurable flow of kernels $K = (K_{s,t})$ will be called *weak* (*generalized*) *solution* of the SDE (5.1) if it satisfies (5.6) without being $\mathcal{F}_{s,t}^W$-measurable.

We have seen that uniqueness is verified if one assumes in addition Wiener measurability: $K_{s,t}$ is $\mathcal{F}_{s,t}^W$-measurable for all $s \le t$.

In case (b) a different consistent system of Feller semigroups $\mathsf{P}_t^{(n),c}$ can be constructed by considering the coalescing $n$-point motion $X_t^{(n),c}$ associated with $X_t^{(n)}$, the $n$-point motion of the Wiener solution. A measurable flow of coalescing maps $\varphi_{s,t}^c$ whose $n$-point motion is $X_t^{(n),c}$ can be defined on an extension $(\Omega, \mathcal{A}, \mathsf{P})$ of the probability space $(\mathcal{W}, \mathcal{F}^W, \mathsf{P}_W)$. This coalescing flow also solves the SDE (5.1). It is a weak solution.

For $s \le t$ set $\mathcal{F}_{s,t}^c = \sigma(\varphi_{u,v}^c,\ s \le u \le v \le t)$. Then $(\mathcal{F}_{s,t}^c)_{s \le t}$ defines a noise. It can be seen (for details see [16]) that $\mathcal{F}_{s,t}^W \subset \mathcal{F}_{s,t}^c$ (this property also holds for any flow solution of SDE). This solution being different from the Wiener solution implies $\mathcal{F}_{s,t}^W \ne \mathcal{F}_{s,t}^c$. The noise $(\mathcal{F}_{s,t}^c)$ cannot be generated by Brownian motions. It is a non-classical noise (see also [27], [28]). The Wiener solution $K^W$ can be recovered by filtering:

$$K_{s,t}^W f = \mathsf{E}[f \circ \varphi_{s,t}^c | \mathcal{F}_{s,t}^W], \quad \text{for all } f \in C(\mathbb{S}).$$

It can be shown that in case (a) there is no weak solution different in law from the Wiener solution. In case (b), $\varphi_{s,t}^c$ is the only solution which is a flow of maps. There are certainly other "intermediate" kernel solutions similar to the sticky flows, but they have not been constructed yet.

**Final remarks.** Similar results hold in a more general context, especially in the case of $\mathbb{S}^d$ and $\mathbb{R}^d$ (including $d = 1$). In fact, for isotropic flows in dimension $d \geq 2$, a different phase appears, in which the Wiener solution is a diffusive flow without hitting. This solution cannot be represented by filtering a coalescing solution defined on an extended probability space and there are no weak (generalized) solutions. In dimension 2 and 3 the coalescing phase (where the Wiener solution is a coalescing flow) and the phase of non uniqueness (where the Wiener solution is diffusive with hitting) still occurs.

Many important questions remain open: for example, the nature of the noises when they are not classical, the possible relations with rough paths ([20]), and the classifications of all solutions, starting with the isotropic case.

## References

[1] Airault, Hélène, and Ren, Jiagang, Modulus of continuity of the canonic Brownian motion "on the group of diffeomorphisms of the circle". *J. Funct. Anal.* **196** (2) (2002), 395–426.

[2] Arratia, R. A., Brownian motion on the line. Ph D Thesis. University of Wisconsin, Madison, 1979.

[3] Baxendale, Peter, Brownian motions in the diffeomorphism group. I. *Compositio Math.* **53** (1) (1984), 19–50.

[4] Bernard, Denis, Gawędzki, Krzysztof, and Kupiainen, Antti, Slow modes in passive advection. *J. Statist. Phys.* **90** (3–4) (1998), 519–569.

[5] Darling, R. W. R., Constructing nonhomeomorphic stochastic flows. *Mem. Amer. Math. Soc.* **70** (376) (1987), vi + 97pp.

[6] E, Weinan, and Vanden Eijnden, Eric, Generalized flows, intrinsic stochasticity, and turbulent transport. *Proc. Natl. Acad. Sci. USA* **97** (15) (2000), 8200–8205 (electronic).

[7] Elworthy, K. D., Stochastic differential equations on manifolds. In *Probability towards* 2000 (New York, 1995), Lecture Notes in Statist. 128, Springer-Verlag, New York 1998, 165–178.

[8] Fang, Shizan, Canonical Brownian motion on the diffeomorphism group of the circle. *J. Funct. Anal.* **196** (1) (2002), 162–179.

[9] Gawedzki, K., and Kupiainen, A., Universality in turbulence: an exactly solvable model. In *Low-dimensional models in statistical physics and quantum field theory* (Schladming, 1995), Lecture Notes in Phys. 469, Springer-Verlag, Berlin 1996, 71–105.

[10] Gawędzki, Krzysztof, and Vergassola, Massimo, Phase transition in the passive scalar advection. *Phys. D* **138** (1-2) (2000), 63–90.

[11] Harris, Theodore E., Coalescing and noncoalescing stochastic flows in $\mathbf{R}^1$. *Stochastic Process. Appl.* **17** (2) (1984), 187–210.

[12] Kunita, Hiroshi, *Stochastic flows and stochastic differential equations*. Cambridge Stud. Adv. Math. 24, Cambridge University Press, Cambridge 1990.

[13] Le Jan, Y., and Lemaire, S., Products of Beta matrices and sticky flows. *Probab. Theory Related Fields* **130** (1) (2004), 109–134.

[14] Le Jan, Yves, On isotropic Brownian motions. *Z. Wahrsch. Verw. Gebiete* **70** (4) (1985), 609–620.

[15] Le Jan, Yves, and Raimond, Olivier, Integration of Brownian vector fields. *Ann. Probab.* **30** (2) (2002), 826–873.

[16] Le Jan, Yves, and Raimond, Olivier, Flows, coalescence and noise. *Ann. Probab.* **32** (2) (2004), 1247–1315.

[17] Le Jan, Yves, and Raimond, Olivier, Sticky flows on the circle and their noises. *Probab. Theory Related Fields* **129** (1) (2004), 63–82.

[18] Le Jan, Yves, and Raimond, Olivier, Flows associated to Tanaka's SDE. *Alea* **1** (2005), 21–34 (electronic).

[19] Le Jan, Yves, and Watanabe, Shinzo, Stochastic flows of diffeomorphisms. In *Stochastic analysis* (Katata/Kyoto, 1982), North-Holland Math. Library 32, North-Holland, Amsterdam 1984, 307–332.

[20] Lyons, Terry, and Qian, Zhongmin, *System control and rough paths*. Oxford Mathematical Monographs, Oxford University Press, Oxford 2002.

[21] Ma, Zhi-Ming, and Xiang, Kai-Nan, Superprocesses of stochastic flows. *Ann. Probab.* **29** (1) (2001), 317–343.

[22] Malliavin, Paul, The canonic diffusion above the diffeomorphism group of the circle. *C. R. Acad. Sci. Paris Sér. I Math.* **329** (4) (1999), 325–329.

[23] Raimond, Olivier, Flots browniens isotropes sur la sphère. *Ann. Inst. H. Poincaré Probab. Statist.* **35** (3) (1999), 313–354.

[24] Tsirelson, Boris, Nonclassical stochastic flows and continuous products. *Probab. Surv.* **1** (2004), 173–298 (electronic).

[25] Tsirelson, Boris, Scaling limit, noise, stability. In *Lectures on probability theory and statistics*, Lecture Notes in Math. 1840, Springer-Verlag, Berlin 2004, 1–106.

[26] Veretennikov, A. Ju., and Krylov, N. V., Explicit formulae for the solutions of stochastic equations. *Mat. Sb.* (*N.S.*) **100** (142) (2) (1976), 266–284, 336.

[27] Warren, J., Splitting: Tanaka's sde revisited. arXiv:math.PR/9911115.

[28] Watanabe, S., The stochastic flow and the noise associated to Tanaka's stochastic differential equation. *Ukraïn. Mat. Zh.* **52** (9) (2000), 1176–1193; English translation *Ukrainian Math. J.* **52** (2) (2000), 1346–1365.

Département Mathématique, Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France

E-mail: Yves.Lejan@math.u-psud.fr

# Stochastic classification models

Peter McCullagh and Jie Yang*

**Abstract.** Two families of stochastic processes are constructed that are intended for use in classification problems where the aim is to classify units or specimens or species on the basis of measured features. The first model is an exchangeable cluster process generated by a standard Dirichlet allocation scheme. The set of classes is not pre-specified, so a new unit may be assigned to a previously unobserved class. The second model, which is more flexible, uses a marked point process as the mechanism generating the units or events, each with its associated class and feature. The conditional distribution given the superposition process is obtained in closed form for one particular marked point process. This distribution determines the conditional class probabilities, and thus the prediction rule for subsequent units.

**Mathematics Subject Classification (2000).** Primary 62H30; Secondary 68T10.

**Keywords.** Cluster process, Cox process, Dirichlet process, Gauss–Ewens process, lack of interference, marked point process, permanent polynomial, Random subset, supervised learning.

## 1. Introduction

**1.1. Classification.** The problem of numerical taxonomy is to classify individual specimens or units $u$ on the basis of measured variables or features $x(u) \in \mathcal{X}$. The units may be anything from tropical insects to bitmap images of handwritten digits or vocalizations of English words. The feature variables may be length or width or weight measurements in the case of insects, or the Fourier transformation at certain frequencies in the case of spoken words. The choice of feature variables is an important problem in its own right, but this matter is of little concern in the present paper.

A deterministic classification model is a rule or algorithm that associates with each feature value $x \in \mathcal{X}$ a class $y(x) \in \mathcal{C}$. Ordinarily the model must be primed or trained on a sample of units with measured features and known classes. In the dialect of artificial intelligence and computer science, the classifier learns the characteristics peculiar to each class and classifies subsequent units accordingly. When the training is over, each subsequent input is a feature value $x(u')$ for a new unit, and the output is the assigned class. The error rate is the fraction of wrong calls.

A stochastic classification model is a process determining a rule that associates with each feature value $x$ a probability distribution $p(\cdot\,; x)$ on the set of classes.

*We are grateful to Jim Pitman for helpful comments. Support for this research was provided by NSF Grant DMS-0305009.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Once again, the classification model must be primed or trained on a sample of units with measured features and known classes. In statistical language, the classifier is a statistical model with unknown parameters to be estimated from the training data. Subsequent units are classified in the usual stochastic sense by computing the conditional distribution given the training data and the feature value for the new unit.

Three stochastic models are described in the sections that follow. The first of these is a regression model with independent components in which the feature values are treated as covariates. The second is an exchangeable cluster process closely related to Fisher's discriminant model, but different in several fundamental ways. The third model is also an exchangeable cluster process, called a permanent cluster process because the conditional distributions are expressed in terms of permanent polynomials.

The distinction between a closed classification model with a pre-determined set of labelled classes, and an open model with unlabelled classes is emphasized. A model of the latter type has a mathematical framework that permits a new unit to be assigned to a class that has not previously been observed and therefore does not have a name. The goal is to construct a classification model with no more than 4–5 parameters to be estimated regardless of the number of classes or the dimension of the feature space. In this way, the technically difficult problems associated with consistency and parameter estimation in high dimensional models are evaded. Ideally, the model should be capable of adapting to classification problems in which one or more classes occupies a non-convex region, or even several disconnected regions, in the feature space.

**1.2. Remarks on the literature.** The literature on stochastic classification is very extensive, the modern theory beginning with Fisher's discriminant model ([12]). Logistic regression models emerged in the 1960s, and with the advent of faster computing, smoothed versions using penalized likelihood became more popular. Stochastic models used in the statistical literature are sometimes complicated, but they are frequently of the most elementary form with independent components such that

$$\log(\mathrm{pr}(Y(u) = r \mid X)) = f_r(X(u)).$$

The goal is to estimate the functions $f_r$ under certain smoothness conditions, which are enforced through penalty functions added to the log likelihood. For a good overview see [29], [15], [27] or [16].

At the more mathematical end of the statistical spectrum, the same model with independent components is frequently used, with $f$ belonging to a suitable space of functions, usually a Besov space. The stated mathematical goal is to obtain the best estimate of $f$ under the most adverse conditions in very large samples ([9]). Smoothing is usually achieved by shrinkage or thresholding of coefficients in a wavelet expansion.

The past decade has seen an upsurge of work in the computer science community under the headings of artificial intelligence, data mining and supervised learning.

Methods used include neural nets, support vector machines and tree classifiers. The emphasis is primarily on algorithms, regularization, efficiency of computation, how best to combine weak classifiers ([13]), and so on. Few algorithms and methods of this type have an overt connection with a generative stochastic process beyond the simple additive form with independent components.

In the Bayesian literature, more complicated processes are constructed using mixture models with Dirichlet priors for the class frequencies ([11], [2], [24], [14]). The cluster process in Section 3 is in fact a simple special case of a more general classification model ([4], [8]). It is used here mainly for illustrative purposes because the distributions can be studied analytically, which is rare for processes generated by Dirichlet allocation schemes.

The semi-parametric models described in Section 4 are of a different type. They are based on Cox processes ([5]) with a baseline intensity measure $\mu$ treated as an unknown parameter. One major attraction for practical work is that the conditional distribution of the class labels given the observed features does not depend on the baseline measure. The unknown nuisance parameter is eliminated by conditioning rather than by integration, and this conditional distribution is the basis for inference and classification.

## 2. Logistic discrimination

**2.1. Non-interference and regression models.** Let $\mathcal{U}$ be the set of units, the infinite set of objects such as plots or subjects or specimens, on which the process $Y$ is defined. A covariate $x\colon \mathcal{U} \to \mathcal{X}$ is a function on the units, the values of which are thought to have an effect on the distribution. In a logistic regression model it is the class $Y(u) \in \mathcal{C}$ that is regarded as the response, and the measured feature $x(u)$ is the covariate.

In practical work, it is often helpful to distinguish between covariates such as sex, age and geographical position that are intrinsic to the unit, and treatment variables such as medication or variety that can in principle be controlled by the experimenter. For mathematical purposes it is more useful to distinguish between a covariate as a function on the units, and a relationship as a function on pairs of units. Examples of the latter include distance if the units are arrayed in space, temporal ordering for time points, genetic or familial relationships if the units are individual organisms, or a block factor as an equivalence relation on units. The statistical distinction, roughly speaking, is that a covariate affects one-dimensional marginal distributions, while a relationship affects bivariate distributions. For present purposes, however, distinctions of this sort are unnecessary.

A regression model is a process in which the joint distribution of the response $(Y(u_1), \ldots, Y(u_n))$ on $n$ units is determined by the covariate values $x = (x(u_1), \ldots, x(u_n))$ on those units. We write $P_n(\cdot\,; x)$ for the joint distribution on an ordered set of $n$ distinct units, implying that two sets of units having the

same ordered list of covariate values, also have the same distribution. In other words, if $(x(u_1), \ldots, x(u_n)) = (x(u'_1), \ldots, x(u'_n))$ then $(Y(u_1), \ldots, Y(u_n))$ and $(Y(u'_1), \ldots, Y(u'_n))$ are both distributed as $P_n(\,\cdot\,; x)$.

In general, the probability assigned to an event $A \subset \mathcal{C}^n$ depends on the covariate vector $(x_1, \ldots, x_n)$. However, the lack of interference condition

$$P_n(A; (x_1, \ldots, x_n)) = P_{n+1}(A \times \mathcal{C}; (x_1, \ldots, x_n, x_{n+1})) \tag{2.1}$$

implies that the probability assigned by $P_{n+1}$ to the event $A \times \mathcal{C}$ does not depend on the final component $x_{n+1}$ of $x$. The failure of this condition means that the probability assigned by $P_2$ to an event of the form $Y(u_1) = 0$ depends on the value of $x(u_2)$. Since the value assigned by $P_1$ to the same event depends only on $x(u_1)$, the two probability distributions are mutually inconsistent. At the very least, interference of this sort may lead to ambiguities in the calculation of probabilities.

Consider two disjoint sets of units with associated vectors $X^{(1)}$, $Y^{(1)}$, $X^{(2)}$, $Y^{(2)}$, all regarded as random variables. Lack of interference is equivalent to the condition that the response $Y^{(1)}$ be conditionally independent of $X^{(2)}$ given $X^{(1)}$. The condition is asymmetric in $X$ and $Y$. As a consequence, the covariate value on unit $u'$ has no effect on the joint distribution for other units. The same term is used in the applied statistical literature ([6], section 2.4; [26]) with a similar meaning, though usually interpreted as a physical or biological property of the system rather than a mathematical property of the model. Without this property, it is difficult to give the model a causal interpretation, so lack of interference is often taken for granted as a logical necessity in applications involving deliberate intervention or assignment of treatment to units.

For applications in which the $x$-values are generated by a process, the preceding argument is not compelling, and the non-interference condition is in fact unduly restrictive. The classification model in Section 3 is derived from an exchangeable bivariate process $(Y(u), X(u))_{u \in \mathcal{U}}$ with finite-dimensional distributions $Q_n$. The conditional distributions $Q_n(\,\cdot \mid X = x)$ determine the joint classification probabilities for $n$ units having the given covariate values as generated by the process. This is not a regression model because the non-interference condition (2.1) is not satisfied by the conditional distributions. As a result, the response distribution for a set of units selected on the basis of their covariate values is not easily determined and is not equal to $Q_n(\,\cdot \mid X = x)$.

We argue that condition (2.1) is unnecessarily strong for certain applications, and that a weaker condition is sufficient for applications in which intervention does not arise. Consider a family of distributions $P_n(\,\cdot\,; x)$, one such distribution for each covariate configuration. It may happen that there exists a bivariate process with distributions $Q_n$ such that, for each covariate configuration $x$ and each event $A \subset \mathcal{C}^n$, the conditional distributions satisfy $P_n(A; x) = Q_n(A \mid X = x)$. The distributions $\{P_n(\,\cdot\,; x)\}$ are then said to be weakly compatible with one another. If such a bivariate process exists, it is not unique because the marginal distribution of the $X$-process is arbitrary. Since the units in the bivariate process have no covariates to distinguish one

from another, the bivariate process is ordinarily exchangeable. Lack of interference implies weak compatibility, but the converse is false.

**2.2. Logistic regression.** In a logistic regression model, the components $Y(u_1), \ldots$ are independent, so the joint distributions are determined by the one-dimensional marginal distributions. The dependence on $x$ is determined by a suitable collection of discriminant functions, $f_j : \mathcal{X} \to \mathcal{R}$, which could be the coordinate projections if $\mathcal{X} = \mathcal{R}^q$, but might include quadratic or other non-linear functions. For a unit $u$ whose feature value is $x = x(u)$ the class probabilities are

$$\log \operatorname{pr}(Y(u) = r) = \sum_j \beta_{rj} f_j(x),$$

where the coefficients $\beta_{rj}$ are parameters to be estimated from the training data. In particular, if there are only two classes, the log odds for class 0 are

$$\log(\operatorname{pr}(Y(u) = 0)/\operatorname{pr}(Y(u) = 1)) = \sum_j (\beta_{0j} - \beta_{1j}) f_j(x). \tag{2.2}$$

For a model with $k$ classes and $q$ linearly independent discriminant functions, the number of parameters is $q(k-1)$, which can be large.

The lack of interference condition is automatically satisfied by the logistic regression model, and in fact by any similar model with independent components.

## 3. An exchangeable cluster process

**3.1. Random permutations and random partitions.** A partition $B$ of the set $[n] = \{1, \ldots, n\}$ is a set of disjoint non-empty subsets called blocks whose union is the whole set. The symbol $\#B$ denotes the number of blocks, and for each block $b \in B$, $\#b$ is the number of elements. The partition is also an equivalence relation on $[n]$, i.e. a function $B : [n] \times [n] \to \{0, 1\}$ that is reflexive, symmetric and transitive. Finally, $B$ is also a symmetric binary matrix with components $B(i, j)$. No distinction is made in the notation between $B$ as a set of subsets, $B$ as a matrix, and $B$ as an equivalence relation. If the partition is regarded as a matrix, $\#B$ is its rank.

Denote by $\mathcal{B}_n$ the set of partitions of $[n]$. Thus, $\mathcal{B}_2 = \{12,\ 1|2\}$ has two elements, and $\mathcal{B}_3$ has five elements

$$123, \quad 12|3, \quad 13|2, \quad 23|1, \quad 1|2|3,$$

where $13|2$ is an abbreviation for $\{\{1, 3\}, \{2\}\}$, containing two blocks. The 15 elements of $\mathcal{B}_4$ can be grouped by block sizes as follows

$$1234, \quad 123|4\ [4], \quad 12|34\ [3], \quad 12|3|4\ [6], \quad 1|2|3|4$$

where 12|34 [3] is an abbreviation for the three distinct partitions 12|34, 13|24, 14|23, each having two blocks of size two. The number of elements in $\mathcal{B}_n$ is the $n$th Bell number, the coefficient of $t^n/n!$ in the generating function $\exp(e^t - 1)$. The first few values are 1, 2, 5, 15, 52, 203, 877,…, increasing rapidly with $n$.

Consider a probability distribution on the symmetric group $\mathcal{S}_n$ in which the probability assigned to the permutation $\sigma$ depends on the number of cycles as follows:

$$p_n(\sigma) = \lambda^{\#\sigma}\Gamma(\lambda)/\Gamma(n+\lambda), \tag{3.1}$$

where $\lambda > 0$, and the ratio of gamma functions is the required normalizing constant. This is the exponential family generated from the uniform distribution with weight function $\lambda^{\#\sigma}$, canonical parameter $\log\lambda$ and canonical statistic $\#\sigma$ the number of cycles. It is evident that the distribution is invariant under the action of the group on itself by conjugation, so $p_n$ is finitely exchangeable. Less obvious but easily verified is the fact that $p_n$ is the marginal distribution of $p_{n+1}$ under the natural deletion operation $\sigma' \mapsto \sigma$ from $\mathcal{S}_{n+1}$ into $\mathcal{S}_n$, which operates as follows. Write $\sigma'$ in cycle form, for example $\sigma' = (1,3)(5)(2,6,4)$ for $n = 5$, and delete element $n+1 = 6$ giving $\sigma = (1,3)(5)(2,4)$. This construction, together with the associated Chinese restaurant process, is described by Pitman ([24], section 4). The projection $\mathcal{S}_{n+1} \to \mathcal{S}_n$ is not a group homomorphism, but successive deletions are commutative. For each $\lambda > 0$, these distributions determine an exchangeable permutation process closely related to the Ewens process on partitions.

The cycles of the permutation $\sigma \in \mathcal{S}_n$ determine a partition of the set $[n]$, and thus a map $\mathcal{S}_n \to \mathcal{B}_n$. The inverse image of $B \in \mathcal{B}_n$ contains $\prod_{b\in B}\Gamma(\#b)$ permutations all having the same probability. Thus, the marginal distribution on partitions induced by (3.1) is

$$p_n(B;\lambda) = \frac{\Gamma(\lambda)\lambda^{\#B}}{\Gamma(n+\lambda)}\prod_{b\in B}\Gamma(\#b) \tag{3.2}$$

for $B \in \mathcal{B}_n$ and $\lambda > 0$ ([10], [1]). This distribution is symmetric in the sense that for each permutation $\sigma\colon [n] \to [n]$, the permuted matrix $(B^\sigma)_{ij} = B_{\sigma(i),\sigma(j)}$ has the same distribution as $B$. The partition $B^\sigma$ has the same block sizes as $B$, which are maximal invariant, and the probability $p_n(B;\lambda)$ depends only on the block sizes. In addition if $B' \sim p_{n+1}(\cdot\,;\lambda)$ is a random partition of $[n+1]$, the leading $n \times n$ submatrix $B$ is a random partition of $[n]$ whose distribution is $p_n(\cdot\,;\lambda)$ ([19]). For each $\lambda > 0$, the sequence of distributions $\{p_n\}$ determines an exchangeable process called the Ewens partition process. For further details, see Pitman ([25]).

The Ewens process is by no means the only example of an exchangeable partition process, but it is one of the simplest and most natural, and it is sufficient to illustrate the ideas in the sections that follow. Some simple extensions are described by Pitman ([24]).

**3.2. Gauss–Ewens cluster process.** A cluster process with state space $\mathcal{X}$ is an infinite sequence of $\mathcal{X}$-valued random variables $X(u)$ for $u \in \mathcal{U}$, together with a random

partition $B\colon \mathcal{U} \times \mathcal{U} \to \{0, 1\}$, which determines the clusters. An observation on a finite set of units $\{u_1, \ldots, u_n\}$ consists of the values $X(u_1), \ldots, X(u_n)$ together with the components of the matrix $B_{ij} = B(u_i, u_j)$. The finite-dimensional distributions on $\mathcal{B}_n \times \mathcal{X}^n$ with densities $p_n$ satisfy the obvious Kolmogorov consistency condition:

$$p_n(B, x_1, \ldots, x_n) = \sum_{B' : \phi B' = B} \int_{\mathcal{X}} p_{n+1}(B', x_1, \ldots, x_{n+1})\, dx_{n+1}$$

where $\phi\colon \mathcal{B}_{n+1} \to \mathcal{B}_n$ is the deletion operator that removes the last row and column.

In the Gauss–Ewens process, $\mathcal{X} = \mathcal{R}^q$ is a vector space. The observation $(B, (X_1, \ldots, X_n))$ on a finite set of $n$ units has a joint density in which $B$ is a partition with distribution (3.2). The conditional distribution given $B$ is Gaussian with constant mean vector, here taken to be zero, and covariance matrix $\Sigma_B = I_n \otimes \Sigma + B \otimes \Sigma_1$, where $\Sigma$, $\Sigma_1$ are $q \times q$ covariance matrices. In component form

$$\operatorname{cov}(X_{ir}, X_{js} \mid B) = \delta_{ij}\Sigma_{rs} + B_{ij}\Sigma_{1\,rs}.$$

This construction implies that $X$ is a sum of two independent processes, one i.i.d. on the units, and one with i.i.d. components for each block.

If $\mathcal{X} = \mathcal{R}$, the coefficient matrices are scalars and the joint density is

$$p_n(B, x) = \frac{\Gamma(\lambda)\lambda^{\#B}}{\Gamma(n + \lambda)} \prod_{b \in B} \Gamma(\#b) \times (2\pi)^{-n/2} |\Sigma_B|^{-1/2} \exp(-x' \Sigma_B^{-1} x/2).$$

It is helpful here to re-parameterize by writing $\bar{x}_b$ for the mean in block $b$, $\theta = \sigma_1^2/\sigma^2$ for the ratio of variance components, $w_b = \#b/(1 + \theta\, \#b)$ and $\bar{x} = \sum w_b \bar{x}_b / \sum w_b$, in which case we have

$$|\Sigma_B|^{-1/2} = \sigma^{-n} \prod_{b \in B} (1 + \theta\, \#b)^{-1/2},$$
$$x' \Sigma_B^{-1} x = \sum_{b \in B} (S^2(b) + w_b \bar{x}_b^2)/\sigma^2,$$

where $S^2(b)$ is the sum of squares for block $b$.

A permutation of the units sends $X_1, \ldots, X_n$ to $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ and also transforms the components of $B$ in such a way that the $i, j$ component of $B^\sigma$ is $B_{\sigma(i)\sigma(j)}$. Evidently, the distribution $p_n$ is unaffected by such permutations, so the Gauss–Ewens process is infinitely exchangeable. As it stands, the Gauss–Ewens process is not a mixture of independent and identically distributed processes because the observation space $\mathcal{B}_n \times \mathcal{X}^n$ for a finite set of $n$ units is not an $n$-fold product space. However, if the blocks are labelled at random, the new process is equivalent in every way to the original, and the new process does follow the de Finetti characterization ([25], p. 44).

**3.3. Conditional distributions.** Given the observed list of feature values $x = (x_1, \ldots, x_n)$, the conditional distribution on partitions induced by the one-dimensional Gauss–Ewens process is

$$p_n(B \mid x) \propto \prod_{b \in B} \lambda\Gamma(\#b)(1 + \theta\,\#b)^{-1/2} \exp((-S^2(b) - w_b\bar{x}_b^2)/(2\sigma^2)).$$

This is a distribution of the product-partition type $p_n(B) \propto \prod_{b\in B} C(b; x)$ ([17]) with cohesion function

$$C(b; x) = \lambda\Gamma(\#b)(1 + \theta\,\#b)^{-1/2} \exp((-S^2(b) - w_b\bar{x}_b^2)/(2\sigma^2))$$

depending on the feature values of the units in block $b$ only. In particular, $C(b; x)$ does not depend on $\#B$ or on $n$. Evidently, two sets of units having the same ordered list of feature values are assigned the same conditional distribution. The marginal distribution on $\mathcal{B}_n$ induced from $p_{n+1}(\cdot \mid (x, x_{n+1}))$ by deleting the last component, depends on the value of $x_{n+1}$, so these conditional distributions do not determine a process. However, there is no contradiction here because these are conditional distributions, and the two conditioning events are different. Since they are derived from a bivariate process, the distributions are weakly compatible with one another in the sense of Section 2.1.

For the multivariate Gauss–Ewens process, the conditional distributions are not of the product-partition type unless the coefficient matrices are proportional, i.e. $\Sigma_1 = \theta\Sigma$. When this condition is satisfied, the cohesion function is an obvious multivariate analogue of the univariate version.

Product partition distributions are certainly convenient for use in applied work, but the great majority of product partition models are incompatible with any process. Consider for example, the product partition model with cohesion function $C(b, x) = \lambda$, independent of the covariate values. For $\lambda = 1$, the distributions are uniform on each $\mathcal{B}_n$. But the distribution on $\mathcal{B}_n$ induced from the uniform distribution on $\mathcal{B}_{n+1}$ is not uniform. The Ewens distributions with cohesion function $\lambda\Gamma(\#b)$ are the only product partition models that are compatible with an exchangeable process.

**3.4. Stochastic classification.** Given the observation $(B, x(u_1), \ldots, x(u_n))$ on $n$ units, plus the feature value $x(u')$ on a subsequent unit, we aim to calculate the conditional distribution $p_{n+1}(\cdot \mid \text{data})$ on $\mathcal{B}_{n+1}$ given the observed values generated by the process. The only missing piece of information is the block to which unit $u'$ is assigned, so the conditional distribution is determined by the probabilities assigned to the events $u' \mapsto b$ for those blocks $b \in B$ or $b = \emptyset$.

A straightforward calculation for a product partition model shows that

$$\text{pr}(u' \mapsto b \mid \text{data}) \propto \begin{cases} C(b \cup \{u'\}, (x, x'))/C(b, x) & b \in B, \\ C(\{u'\}, x') & b = \emptyset, \end{cases}$$

where $(x, x')$ is the complete list of $n+1$ observed feature values. For $b \in B$, the cohesion ratio for the univariate Gauss–Ewens process is

$$\#b\gamma^{1/2} \exp(-\gamma(x' - \theta\#b\bar{x}_b/(1+\theta\#b))^2/(2\sigma^2))$$

where $\gamma = (1+\theta\#b)/(1+\theta(\#b+1))$. If $\theta\#b$ is large, blocks whose sample means are close to $x'$ have relatively high probability, which is to be expected.

The predictive distribution for the general multivariate Gauss–Ewens process involves a ratio of multivariate normal densities. Although preference is given to larger blocks, the predictive distribution also puts more weight on those classes whose block means are close to $x'$. If $x'$ is sufficiently far removed from all observed block means, the empty set (new class) is given relatively greater weight. When the empty set is excluded from consideration the parameter $\lambda$ has no effect, and the predictive distribution is roughly the same as that obtained from the Fisher discriminant model with prior probabilities proportional to class sizes.

## 4. Point process models

**4.1. Permanent polynomial.** To each square matrix $K$ of order $n$ there corresponds a polynomial of degree $n$,

$$\mathrm{per}_t(K) = \sum_\sigma t^{\#\sigma} K_{1\sigma(1)} \cdots K_{n\sigma(n)}$$

where the sum runs over permutations of $\{1, \ldots, n\}$, and $\#\sigma$ is the number of cycles. The conventional permanent is the value at $t = 1$, and the determinant is $\det(K) = \mathrm{per}_{-1}(-K)$. The coefficient of $t$ is the sum of cyclic products

$$\mathrm{cyp}(K) = \lim_{t\to 0} t^{-1} \mathrm{per}_t(K) = \sum_{\sigma:\#\sigma=1} K_{1\sigma(1)} \cdots K_{n\sigma(n)}.$$

For certain types of patterned matrices, the permanent polynomial can be evaluated in closed form or by recursion. Consider, for example, the matrix $J$ of order $n$ such that $J_{ii} = \zeta$ and $J_{ij} = 1$ otherwise. The permanent polynomial is the value $f_n(t)$ obtained by recursion

$$\begin{pmatrix} f_{n+1}(t) \\ h_{n+1}(t) \end{pmatrix} = \begin{pmatrix} \zeta t & n \\ t & n \end{pmatrix} \begin{pmatrix} f_n(t) \\ h_n(t) \end{pmatrix}$$

starting with $f_0(t) = h_0(t) = 1$. In particular, for $\zeta = 1$ and $t = \lambda$ we obtain the value $f_n(\lambda) = \Gamma(n+\lambda)/\Gamma(\lambda)$, which is the normalizing constant in the distribution (3.1).

**4.2. Gaussian moments.** The permanent polynomial arises naturally in statistical work associated with factorial moment measures of Cox processes as follows. Let $Z$ be

a zero-mean real Gaussian process on $\mathcal{X}$ with covariance function $\operatorname{cov}(Z(x), Z(x')) = K(x, x')/2$. The joint cumulant and the joint moment of the squared variables $|Z(x_1)|^2, \ldots, |Z(x_n)|^2$ are

$$\operatorname{cum}_n\big(|Z(x_1)|^2, \ldots, |Z(x_n)|^2\big) = \operatorname{cyp}[K](x_1, \ldots, x_n)/2,$$
$$E\big(|Z(x_1)|^2 \cdots |Z(x_n)|^2\big) = \operatorname{per}_{1/2}[K](x_1, \ldots, x_n),$$

where $[K](x_1, \ldots, x_n)$ is the symmetric matrix of order $n$ whose entries are $K(x_i, x_j)$. More generally, if $\Lambda(x) = |Z_1(x)|^2 + \cdots + |Z_k(x)|^2$ is the sum of squares of $k$ independent and identically distributed Gaussian processes, we have

$$\begin{aligned} \operatorname{cum}_n(\Lambda(x_1), \ldots, \Lambda(x_n)) &= \alpha \ \operatorname{cyp}[K](x_1, \ldots, x_n), \\ E(\Lambda(x_1) \cdots \Lambda(x_n)) &= \operatorname{per}_\alpha[K](x_1, \ldots, x_n) \end{aligned} \tag{4.1}$$

with $\alpha = k/2$ ([22]). Thus, if $\Lambda$ is the intensity function for a doubly stochastic Poisson process, the $n$th order product density at $x = (x_1, \ldots, x_n)$ is $\operatorname{per}_\alpha[K](x)$. In other words, the expected number of ordered $n$-tuples of distinct events occurring in an infinitesimal ball of volume $dx$ centered at $x \in \mathcal{X}^n$ is $\operatorname{per}_\alpha[K](x)\, dx$.

The analogous result for zero-mean complex-valued processes with covariance function $\operatorname{cov}(Z(x), \bar{Z}(x')) = K(x, x')$ and $\Lambda$ as defined above is the same except that $\alpha = k$ rather than $k/2$. A proof for $\alpha = 1$ can be found in Macchi ([21]), and for general $k$ in McCullagh and Møller ([22]). Although $K$ is Hermitian, the polynomial is real because inverse permutations have conjugate coefficients.

**4.3. Convolution semi-group properties.** Permanent polynomials also have a semi-group convolution property that is relevant for probability calculations connected with the superposition of independent processes. In describing this property, it is helpful to regard the points $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ as distinct and unordered, so $\boldsymbol{x}$ is a finite subset of $\mathcal{X}$. Since $\operatorname{per}_\alpha[K](x_1, \ldots, x_n)$ is a symmetric function of $x$, we may write $\operatorname{per}_\alpha[K](\boldsymbol{x})$ without ambiguity for non-empty sets. For the empty subset, $\operatorname{per}_\alpha[K](\emptyset) = 1$. It is shown in McCullagh and Møller ([22]) that

$$\sum_{\boldsymbol{w} \subset \boldsymbol{x}} \operatorname{per}_\alpha[K](\boldsymbol{w}) \operatorname{per}_{\alpha'}[K](\bar{\boldsymbol{w}}) = \operatorname{per}_{\alpha+\alpha'}[K](\boldsymbol{x}) \tag{4.2}$$

where the sum is over all $2^n$ subsets, and $\bar{\boldsymbol{w}}$ is the complement of $\boldsymbol{w}$ in $\boldsymbol{x}$.

Suppose that $\operatorname{per}_\alpha[K](\boldsymbol{x})$ is the density at $\boldsymbol{x}$, with respect to some product measure $\mu(dx_1) \cdots \mu(dx_n)$, of a finite point process in $\mathcal{X}$. The convolution property implies that the superposition of two independent processes having the same covariance function $K$ has a distribution in the same family with parameter $\alpha + \alpha'$. Furthermore, the ratio

$$q(\boldsymbol{w}; \boldsymbol{x}) = \frac{\operatorname{per}_\alpha[K](\boldsymbol{w}) \operatorname{per}_{\alpha'}[K](\bar{\boldsymbol{w}})}{\operatorname{per}_{\alpha+\alpha'}[K](\boldsymbol{x})} \tag{4.3}$$

determines a probability distribution on the subsets of $\boldsymbol{x}$. If in fact some components of $\boldsymbol{x}$ are duplicated, these duplicates must be regarded as distinct units that happen to have the same $x$-value, and $q$ is then regarded as a distribution on subsets of the $n$ units. In the extreme case where all components are identical, all components of the matrix $[K](\boldsymbol{x})$ are equal, and the distribution reduces to

$$q(\boldsymbol{w};\boldsymbol{x}) = \frac{\Gamma(\#\boldsymbol{w}+\alpha)\,\Gamma(\#\overline{\boldsymbol{w}}+\alpha')\,\Gamma(\alpha+\alpha')}{\Gamma(n+\alpha+\alpha')\,\Gamma(\alpha)\,\Gamma(\alpha')}.$$

In other words, $\#\boldsymbol{w}$ has the beta-binomial distribution.

The statistical construction ensures that the polynomial $\mathrm{per}_\alpha(K)$ is positive at all positive half-integer values of $\alpha$ provided only that $K$ is real symmetric and positive semi-definite. In view of the convolution property, it is natural to ask whether the permanent polynomial of a real symmetric positive semi-definite matrix is positive for all $\alpha \geq 1/2$. The numerical evidence on this point is compelling, but so far there is no proof. On the one hand, there exist positive semi-definite symmetric matrices such that $\mathrm{per}_\alpha(K) < 0$ for values in the interval $0 < \alpha < 1/2$. On the other hand, extensive numerical work has failed to produce a positive semi-definite matrix such that the permanent polynomial has a root whose real part exceeds one half. Although no proof is offered, it seems safe to proceed as if $\mathrm{per}_\alpha(K) \geq 0$ for all $\alpha \geq 1/2$ and positive semi-definite symmetric $K$. In applications where the covariance function is non-negative, the permanent polynomial is clearly positive for all $\alpha > 0$.

**4.4. A marked point process.** Consider a Poisson process $\boldsymbol{X}$ in $\mathcal{X}$ with intensity measure $\mu$. In the first instance, $\boldsymbol{X}$ is a counting measure in $\mathcal{X}$ such that the number of events $\boldsymbol{X}(A)$ has the Poisson distribution with mean $\mu(A)$. In addition, for non-overlapping sets $A, A'$, the event counts $\boldsymbol{X}(A)$ and $\boldsymbol{X}(A')$ are independent. The process is said to be regular if it has no multiple events at the same point and is finite on compact sets. In that case $\boldsymbol{X}$ is a random subset of $\mathcal{X}$ such that $\boldsymbol{X} \cap A$ is finite for compact sets $A$. For linguistic convenience, we use the terminology associated with random sets rather than the terminology associated with random measures or multisets. All processes are assumed to be regular.

A Poisson process driven by a random intensity measure $\Lambda(x)\mu(dx)$ is called a doubly stochastic Poisson process, or a Cox process. Details of such processes can be found in the books by Kingman ([20]) and Daley and Vere-Jones ([7]).

Let $\mu$ be a non-random measure in $\mathcal{X}$ serving as a baseline for the construction of subsequent point processes. For probabilistic purposes, $\mu$ is a fixed measure defined on a suitable algebra of subsets of $\mathcal{X}$ that includes all singletons. For statistical purposes, $\mu$ is a parameter to be estimated, if necessary, from the data. Given a random non-negative intensity function $\Lambda(x)$, the associated Cox process is such that the expected number of events occurring in an infinitesimal ball $dx$ centered at $x$ is $E(\Lambda(x))\,\mu(dx)$. Likewise, the expected number of ordered pairs of distinct events in the infinitesimal product set $dx\,dx'$ at $(x, x')$ is $E(\Lambda(x)\Lambda(x'))\,\mu(dx)\,\mu(dx')$, and

so on. In general, for $\boldsymbol{x} = (x_1, \ldots, x_n)$,

$$m^{(n)}(\boldsymbol{x}) = E(\Lambda(x_1) \cdots \Lambda(x_n))$$

is called the $n$th order product density at $\boldsymbol{x} \in \mathcal{X}^n$. These expectations are the densities of the factorial moment measures of the process with respect to the product measure $\mu^n$. The order is implicit from the argument $\boldsymbol{x} \in \mathcal{X}^n$, so we usually write $m(\boldsymbol{x})$ rather than $m^{(n)}(\boldsymbol{x})$.

Ordinarily, in typical ecological applications or studies of the spatial interactions of particles, an observation on a point process consists of a census $\boldsymbol{X} \cap S$ of all events occurring in the bounded set $S$. The observation tells us not only that an event occurred at certain points in $S$, but also that no events occurred elsewhere in $S$. For the sorts of applications with which we are concerned, however, the training sample is not exhaustive, so the observation is regarded as a sample of the events in $\mathcal{X}$. Such an observation tells us only that an event occurred at certain points in $\mathcal{X}$, and says nothing about the occurrence or non-occurrence of events elsewhere.

Suppose now that $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}$ are $k$ independent Cox process on $\mathcal{X}$ driven by independent random intensity functions $\Lambda_1(x), \ldots, \Lambda_k(x)$, all relative to the same measure $\mu$. The marked process can be represented by the pair $(\boldsymbol{X}, y)$ in which $\boldsymbol{X} = \cup \boldsymbol{X}^{(r)}$ is the superposition process, and $y\colon \boldsymbol{X} \to \mathcal{C}$ is the list of labels. Then the $r$th component process $\boldsymbol{X}^{(r)} = y^{-1}(r)$ is the inverse image of label $r$.

Let $\boldsymbol{x} \subset \mathcal{X}$ be a given finite point configuration consisting of $n$ points. Given that $\boldsymbol{x} \subset \boldsymbol{X}$, i.e. that the superposition process contains $\boldsymbol{x}$, each event $x \in \boldsymbol{x}$ has a label $y(x)$ in the marked process so there are $k^n$ possible values for the labels of the events in $\boldsymbol{x}$. Denote by $\boldsymbol{x}^{(r)}$ the subset $\boldsymbol{x} \cap y^{-1}(r)$, possibly empty, consisting of those events in $\boldsymbol{x}$ having label $r$. The conditional distribution of the class labels given $\boldsymbol{x} \subset \boldsymbol{X}$ is proportional to the product of the product densities of the component processes

$$p_n(y \mid \boldsymbol{x}) = \frac{\prod_{r \in \mathcal{C}} m_r(\boldsymbol{x}^{(r)})}{m_\cdot(\boldsymbol{x})}. \tag{4.4}$$

In this expression, $m_r(\boldsymbol{x}^{(r)})$ is the product density of order $\#\boldsymbol{x}^{(r)}$ at $\boldsymbol{x}^{(r)}$ for the process labelled $r$, and $m_\cdot(\boldsymbol{x})$ is the $n$th order product density for the superposition process at $\boldsymbol{x}$. For the empty set, $m_r(\emptyset) = 1$. A key point to note is that the conditional distribution of the class labels depends only on the product densities, and not on the baseline measure $\mu$.

The conditional distribution of the unlabelled partition $B$ is obtained by ignoring labels, in effect by multiplying by the combinatorial coefficient $k!/(k - \#B)!$. Since the combinatorial coefficient depends on the number of blocks, the conditional distribution of the unlabelled partition is not a product partition model, but it is a distribution of Gibbs type ([25], p. 26)

These conditional distributions do not determine a regression model because they fail to satisfy the lack of interference condition (2.1). However, they are derived from a bona fide bivariate process, so they are mutually compatible in the weak sense.

In this context of prediction, it may be helpful to think of each event as a unit or specimen, in such a way that $x(u)$ is the position or feature value of the event, and $y(u)$ is the label. To classify a new unit or event $u'$ such that $x(u') = x'$, it is sufficient to calculate the conditional distribution as determined by $p_{n+1}$ given the extended configuration $\boldsymbol{x}' = \boldsymbol{x} \cup \{x'\}$ plus the labels of those points in $\boldsymbol{x}$. The conditional probabilities are proportional to the ratio of product densities

$$p_{n+1}(y(u') = r \mid \text{data}) \propto m_r(\boldsymbol{x}^{(r)} \cup \{x'\})/m_r(\boldsymbol{x}^{(r)}) \tag{4.5}$$

for $r \in \mathcal{C}$.

**4.5. Specific examples.** We consider two examples, one in which the intensity is the square of a Gaussian process with product density (4.1), and one in which the intensity is log normal.

*Permanent process.* Suppose that each component process is a permanent process and that the product density for process $r$ is $m_r(\boldsymbol{x}) = \text{per}_{\alpha_r}[K](\boldsymbol{x})$. Then the product density for the superposition process is $\text{per}_{\alpha_\cdot}[K](\boldsymbol{x})$ and the conditional distribution of the labels given $\boldsymbol{x}$ is

$$p_n(y \mid \boldsymbol{x}) = \frac{\text{per}_{\alpha_1}[K](\boldsymbol{x}^{(1)}) \cdots \text{per}_{\alpha_k}[K](\boldsymbol{x}^{(k)})}{\text{per}_{\alpha_\cdot}[K](\boldsymbol{x})}. \tag{4.6}$$

This distribution determines a random labelled partition of the given events into $k$ classes, some of which may be empty. It is the 'multinomial' generalization of (4.3), and is closed under aggregation of classes.

For a new unit $u'$ such that $x(u') = x'$, the conditional probability of class $r$ is proportional to the permanent ratio

$$p_{n+1}(y(u') = r \mid \text{data}) \propto \text{per}_{\alpha_r}[K](\boldsymbol{x}^{(r)}, x')/\,\text{per}_{\alpha_r}[K](\boldsymbol{x}^{(r)}).$$

This expression is restricted to the set of $k$ classes in $\mathcal{C}$, but it may include classes for which $\boldsymbol{x}^{(r)}$ is empty, i.e. named classes that do not occur in the training sample. In the extreme case where $\boldsymbol{x}$ is empty, the probability of class $r$ is $\alpha_r/\alpha_\cdot$ regardless of $x'$.

The derivation of the conditional distribution from the marked point process requires each $\alpha$ to be a half-integer, and $K$ to be positive semi-definite. Alternatively, $K$ could be Hermitian and $\alpha_r$ a whole integer. However, if $K$ is non-negative on $\mathcal{X}$, the distribution (4.6) exists for arbitrary $\alpha_r > 0$, even if $K$ is not positive semi-definite. We shall therefore consider the limit in which $\alpha_r = \alpha$ and $k \to \infty$ such that $\alpha_\cdot = k\alpha = \lambda > 0$ is held fixed. The limit distribution for the unlabelled partition is

$$p_n(B \mid \boldsymbol{x}; \lambda) = \frac{\lambda^{\#B} \prod_{b \in B} \text{cyp}[K](\boldsymbol{x}^{(b)})}{\text{per}_\lambda[K](\boldsymbol{x})}, \tag{4.7}$$

which is a product partition model, and reduces to the Ewens distribution if $K$ is constant on $\mathcal{X}$. For a new unit $u'$ such that $x(u') = x'$, the conditional probability of

assignment to block $b$ is

$$p_{n+1}(u' \mapsto b \mid \text{data}) \propto \begin{cases} \text{cyp}[K](\boldsymbol{x}^{(b)}, x')/\,\text{cyp}[K](\boldsymbol{x}^{(b)}) & b \in B, \\ \lambda K(x', x') & b = \emptyset. \end{cases}$$

Our experience with these classification rules is restricted to the simplest versions of the model in which $\mathcal{X}$ is Euclidean space and $K(x, x') = \exp(-|x - x'|^2/\rho^2)$ or similar versions such as $\exp(-|x - x'|/\rho)$. On the whole, the smoother version is better, and the value of $\alpha$ in (4.6) has only minor effects. It is necessary to select a suitable value of the range parameter $\rho$, but the qualitative conclusions are the same for all $\rho$. The region in the $\mathcal{X}$-space for which the predictive probability of class $r$ is high need not be convex or simply connected. In that sense, both of these classification rules are qualitatively different from the one derived from the Gauss–Ewens process.

*Log Gaussian Cox processes.* Suppose that each component process is log Gaussian, i.e. $\log \Lambda_r$ is a Gaussian process with mean and variance

$$E \log \Lambda_r(x) = \theta_r(x), \quad \text{cov}(\log \Lambda_r(x), \log \Lambda_r(x')) = K_r(x, x').$$

Then the $n$th order product density at $x = (x_1, \ldots, x_n)$ is

$$m_r(x) = \exp\Big(\sum_j \theta_r(x_j) + \tfrac{1}{2} \sum_{ij} K_r(x_i, x_j)\Big).$$

Given that $\boldsymbol{x}$ occurs in the superposition process, the conditional distribution of the labels satisfies

$$\log p_n(y \mid \boldsymbol{x}) = \sum_{x \in \boldsymbol{x}} \theta_{y(x)}(x) + \tfrac{1}{2} \sum_{\substack{x, x' \in \boldsymbol{x} \\ y(x) = y(x')}} K_{y(x)}(x, x') + \text{const.}$$

Finally, a new unit with $x(u') = x'$ generated from the process is assigned to class $r$ with probability

$$\log p_{n+1}(y(u') = r \mid \text{data}) = \theta_r(x') + \tfrac{1}{2} K_r(x', x') + \sum_{x \in \boldsymbol{x}^{(r)}} K_r(x', x) + \text{const.}$$

Thus, if $\theta_r(x) = \sum_j \beta_{rj} f_j(x)$ as in Section 2.2, and there are only two classes with $K_0 = K_1 = K$, the conditional log odds that the new unit is assigned to class 0 are

$$\sum_j (\beta_{0j} - \beta_{1j}) f_j(x') + \sum_{x \in \boldsymbol{x}^{(0)}} K(x', x) - \sum_{x \in \boldsymbol{x}^{(1)}} K(x', x), \tag{4.8}$$

coinciding with (2.4) when $K = 0$.

**4.6. Numerical illustration.** A simple artificial example suffices to illustrate the qualitative difference between classification models based on Cox processes, and classification models of the type described in Section 3. We use the two-class permanent model (4.6) with $\alpha_1 = \alpha_2 = 1$. The feature space is a $3 \times 3$ square in the plane, the covariance function is $K(x, x') = \exp(-\|x - x'\|^2/\rho^2)$ with $\rho = 0.5$, and the true class is determined by a $3 \times 3$ chequerboard pattern with white in the center square. The training data consists of 90 units, with 10 feature values uniformly distributed in each small square as shown in the first panel of Figure 1. The second panel is a density plot, and the third panel a contour plot, of the conditional probability that a new unit at that point is assigned to class 'white'. These probabilities were computed by an approximation using a cycle expansion for the permanent ratio.

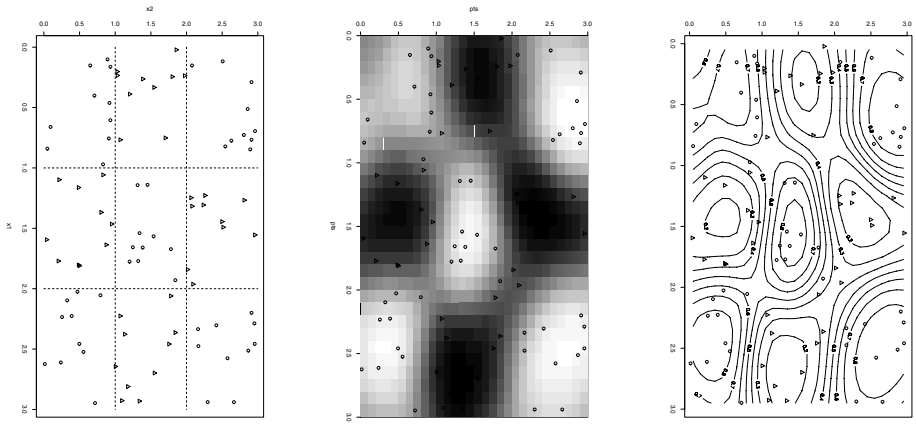


Figure 1. Predictive probability of class I using a permanent model.

For the parameter values chosen, the range of predictive probabilities depends to a moderate extent on the configuration of $x$-values in the training sample, but the extremes are seldom below 0.1 or above 0.9 for a configuration of 90 points with 10 in each small square. The range of predictive probabilities decreases as $\rho$ increases, but the 50% contour is little affected, so the classification is fairly stable. Given that the correct classification is determined by the chequerboard rule, the error rate for the permanent model using this particular training configuration can be computed exactly: it is around 13% for a point chosen uniformly at random from the large square. This error rate is a little misleading because most of those errors occur near an internal boundary where the predictive probability is close to 0.5. Gross errors are rare.

## 5. Parameter estimation

Let $(y, \boldsymbol{x})$ be the training data, and let $x'$ be the feature value for a subsequent unit. In principle, the likelihood function should be computed for the full data including the

value for the subsequent unit. In practice, it is more convenient to base the likelihood on the training data alone, i.e. $p_n(y, \boldsymbol{x}; \theta)$ at the parameter point $\theta$. Ordinarily, the information sacrificed by ignoring the additional factor is negligible for large $n$, and the gain in simplicity may be substantial.

Likelihood computations are straightforward for logistic regression models, and the same is true for the Gauss–Ewens process, but the state of affairs is more complicated for point process models. Consider a marked permanent process model with $\alpha_r = \alpha$, in which $\mathcal{X}$ is a Euclidean space and $K(x, x') = \exp(-\|x - x'\|^2/\rho^2)$. The parameters of the process are the scalars $\alpha$, $\rho$ plus the baseline measure $\mu$. However, the conditional likelihood given the observation $\boldsymbol{x}$ from the training sample depends only on $\alpha$, $\rho$, and the predictive distribution also depends only on $(\alpha, \rho)$. In this setting, the distribution of $\boldsymbol{x}$ is governed largely by the baseline measure $\mu$, so the information for $(\alpha, \rho)$ in the superposition process must be negligible. Accordingly, we use the conditional likelihood instead of the full likelihood, for parameter estimation.

Even though the most troublesome component of the parameter has been eliminated, computation of the likelihood for the remaining parameters does present difficulties. In the case of the log Gaussian model, the normalizing constant is not available in closed form. In the case of the permanent models (4.6) or (4.7), for which the normalizing constants are available, the only remaining obstacle is the calculation of cyclic products and permanent polynomials. The permanent of a large matrix is notoriously difficult to compute exactly ([28]), and the permanent polynomial appears to be even more challenging. For $\alpha = 1$, polynomial-time algorithms are available for fixed-rank matrices ([3]). In addition, the existence of polynomial-time Monte Carlo algorithms for non-negative matrices, has been demonstrated but not implemented ([18]).

Our experience for positive definite matrices is less pessimistic than the preceding remarks suggest. Reasonably accurate polynomial-time continued-fraction approximations for the ratio of permanent polynomials can be developed without resorting to Monte Carlo approximation. We use a cycle expansion whose accuracy improves as $\alpha$ increases. Here, reasonably accurate means within 2–3% for typical covariance matrices of order $n = 100$, and for $\alpha \geq 1/2$. These expansions, which were used in the construction of Figure 1, will be described elsewhere.

## References

[1] Aldous, D., Probability distributions on cladograms. In *Random Discrete Structures* (ed. by D. Aldous and R. Pemantle), IMA Vol. Math. Appl. 76, Springer-Verlag, New York 1995, 1–18.

[2] Antoniak, C. E., Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** (1974), 1152–1174.

[3] Barvinok, A. I., Two algorithmic results for the traveling salesman problem. *Math. Oper. Res.* **21** (1996), 65–84.

[4] Blei, D., Ng, A., Jordan, M., Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** (2003), 993–1022.

[5] Cox, D. R., Some statistical methods connected with series of events. *J. Roy. Statist. Soc. Ser. B* **17** (1955), 129–164.

[6] Cox, D. R., *Planning of Experiments*. Wiley Publ. Appl. Statist., Wiley, New York 1958.

[7] Daley D., Vere-Jones, D., *An Introduction to the Theory of Point Processes*. 2nd edition, Probab. Appl. (N. Y.), Springer-Verlag, New York 2003.

[8] Daumé, H., Marcu, D., A Bayesian model for supervised clustering with the Dirichlet process prior. *J. Mach. Learn. Res.* **6** (2005), 1551–1577.

[9] Donoho, D., Johnstone, I., Kerkyacharian, G., Picard, D., Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** (1995), 301–369.

[10] Ewens, W. J., The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3** (1972), 87–112.

[11] Ferguson, T., A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** (1973), 209–230.

[12] Fisher, R. A., The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (1936), 179–188.

[13] Freund, Y., Schapire, R., Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kauffman, San Francisco 1996, 148–156.

[14] Gopalan, R., Berry, D., Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93** (1998), 1130–1139.

[15] Green, P., Silverman, B., *Nonparametric Regression and Generalized Linear Models*. Monogr. Statist. Appl. Probab. 58, Chapman and Hall, London, 1994.

[16] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Ser. Statist., Springer-Verlag, New York 2001.

[17] Hartigan, J. A., Partition models. *Comm. Statist. Theory Methods* **19** (1990), 2745–2756.

[18] Jerrum, M., Sinclair, A., Vigoda, E., A polynomial-time approximation algorithm for approximating the permanent of a matrix with non-negative entries. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2001, 712–721.

[19] Kingman, J. F. C., *Mathematics of Genetic Diversity*. CBMS-NSF Regional Conf. Ser. in Appl. Math. 34, SIAM, Philadelphia, PA, 1980.

[20] Kingman, J. F. C., *Poisson Processes*. Oxford Stud. Probab. 3, Clarendon Press, Oxford University Press, Oxford 1993.

[21] Macchi, O., The coincidence approach to stochastic point processes. *Adv. in Appl. Probab.* **7** (1975), 83–122.

[22] McCullagh, P., Møller, J., The permanent process. 2005; available via http://www.stat.uchicago.edu/~pmcc/permanent.pdf.

[23] Minc, H., *Permanents*. Encyclopedia Math. Appl. 6, Addison-Wesley, Reading, MA, 1978.

[24] Pitman, J., Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (ed. by T. S. Ferguson et al.), IMS Lecture Notes Monogr. Ser. 30, Hayward, CA, 1996, 245–267.

[25] Pitman, J., *Combinatorial Stochastic Processes* (Ecole d'Eté de Probabilités de Saint-Flour XXXII, 2002). Lecture Notes in Math. 1875, Springer-Verlag, Berlin 2006.

[26] Rubin, D., Which Ifs have causal answers? *J. Amer. Statist. Assoc.* **81** (1986), 961–962.

[27] Ripley, B., *Pattern recognition and Neural Networks*. Cambridge University Press, Cambridge 1996.

[28] Valiant, L. G., The complexity of computing the permanent. *Theoret. Comput. Sci.* **8** (1979), 189–201.

[29] Wahba, G., *Spline Models for Observational Data*. CBMS-NSF Regional Conf. Ser. in Appl. Math. 59. SIAM, Philadelphia, PA, 1990.

Department of Statistics, University of Chicago, 5734 S. University Ave, Chicago, IL 60637, U.S.A.

E-mail: pmcc@galton.uchicago.edu

Department of Statistics, University of Chicago, 5734 S. University Avenue, Eckhart 108, Chicago, IL 60637 U.S.A.

E-mail: jyang@galton.uchicago.edu

# Random partitions and instanton counting

Andrei Okounkov*

**Abstract.** We summarize the connection between random partitions and $\mathcal{N} = 2$ supersymmetric gauge theories in 4 dimensions and indicate how this relation extends to higher dimensions.

**Mathematics Subject Classification (2000).** Primary 81T13; Secondary 14J60.

## 1. Introduction

**1.1. Random partitions.** A partition of $n$ is a monotone sequence

$$\lambda = (\lambda_1 \geq \lambda_2 \geq \cdots \geq 0)$$

of nonnegative integers with sum $n$. The number $n$ is denoted $|\lambda|$ and called the size of $\lambda$. A geometric object associated to a partition is its *diagram*; it contains $\lambda_1$ squares in the first row, $\lambda_2$ squares in the second row and so on. An example, flipped and rotated by $135^\circ$ can be seen in Figure 1. Partitions naturally label many basic objects

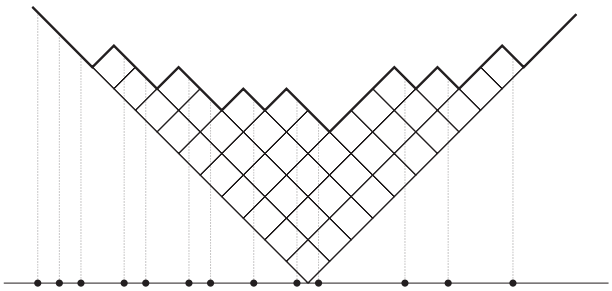


Figure 1. The diagram of $\lambda = (10, 8, 7, 4, 4, 3, 2, 2, 1, 1)$, flipped and rotated by $135^\circ$. Bullets indicate the points of $\mathfrak{S}(\lambda)$. The profile of $\lambda$ is plotted in bold.

in mathematics and physics, such as e.g. conjugacy classes and representations of the symmetric group $S(n)$, and very often appear as modest summation ranges or indices. A simple but fruitful change of perspective, which I wish to stress here, is to treat sums over partitions probabilistically, that is, treat them as expectations of some functions of a random partition.

*The author thanks Packard Foundation for partial financial support.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

A survey of the theory of random partitions may be found in [33]. Of the several natural measures on partitions, the Plancherel measure

$$\mathfrak{M}_{\mathrm{Planch}}(\lambda) = \frac{(\dim \lambda)^2}{n!}, \quad |\lambda| = n, \tag{1}$$

stands out as the one with deepest properties and widest applications. Here $\dim \lambda$ is the dimension of the corresponding representation of $S(n)$. This is a probability measure on partitions of $n$. It can be viewed as a distinguished discretization of the GUE ensemble of the random matrix theory. Namely, a measure on partitions can be made a point process on a lattice by placing particles in positions

$$\mathfrak{S}(\lambda) = \left\{\lambda_i - i + \tfrac{1}{2}\right\} \subset \mathbb{Z} + \tfrac{1}{2}.$$

Figure 1 illustrates the geometric meaning of this transformation. An important theme of recent research was to understand how and why for a Plancherel random partition of $n \to \infty$ the particles $\mathfrak{S}(\lambda)$ behave like the eigenvalues of a random Hermitian matrix. See [3], [4], [14], [31] and e.g. [15], [32], [33] for a survey.

In these notes, we consider a different problem, namely, the behavior of (1) in a *periodic potential*, that is, additionally weighted by a multiplicative periodic function of the particles' positions. This leads to new phenomena and new applications. As we will see, the partition function of $\mathfrak{M}_{\mathrm{Planch}}$ in a periodic potential is closely related to Nekrasov partition function from supersymmetric gauge theory. This relationship will be reviewed in detail in Section 2 and its consequences may be summarized as follows.

**1.2. Instanton counting.** In 1994, Seiberg and Witten proposed an exact description of the low-energy behavior of certain supersymmetric gauge theories [38], [39]. In spite of the enormous body of research that this insight has generated, only a modest progress was made towards its gauge-theoretic derivation. This changed in 2002, when Nekrasov proposed in [28] a physically meaningful and mathematically rigorous definition of the regularized partition function $Z$ for supersymmetric gauge theories in question.

Supersymmetry makes the gauge theory partition function the partition function of a *gas of instantons*. Nekrasov's idea was to use *equivariant integration* with respect to the natural symmetry group in lieu of a long-distance cut-off for the instanton gas. He conjectured that as the regularization parameter $\varepsilon \to 0$

$$\ln Z \sim -\frac{1}{\varepsilon^2}\mathcal{F}$$

where the free energy $\mathcal{F}$ expressed by the Seiberg–Witten formula in terms of periods of a certain differential $dS$ on a certain algebraic curve $C$.

This conjecture was proven in 2003 by Nekrasov and the author for a list of gauge theories with gauge group $U(r)$, namely, pure gauge theory, theories with matter

fields in fundamental and adjoint representations of the gauge group, as well as 5-dimensional theory compactified on a circle [29]. Simultaneously, independently, and using completely different ideas, the formal power series version of Nekrasov's conjecture was proven for the pure $U(r)$-theory by Nakajima and Yoshioka [25]. The methods of [29] were applied to classical gauge groups in [30] and to the 6-dimensional gauge theory compactified on a torus in [11]. Another algebraic approach, which works for pure gauge theory with any gauge group, was developed by Braverman [5] and Braverman and Etingof [6].

In these notes, we outline the results of [29] in the simplest, yet fundamental, case of pure gauge theory. As should be obvious from the title, the main idea is to treat the gauge theory partition function $Z$ as the partition function of an ensemble of random partitions. The $\varepsilon \to 0$ limit turns out to be the *thermodynamic limit* in this ensemble. What emerges in this limit is a nonrandom *limit shape*, an example of which may be seen in Figure 6. This is a form of the law of large numbers, analogous, for example, to Wigner's semicircle law for the spectrum of a large random matrix. The limit shape is characterized as the unique minimizer $\psi_\star$ of a certain convex functional $\mathcal{S}(\psi)$, leading to

$$\mathcal{F} = \min \mathcal{S}.$$

We solve the variational problem explicitly and the limit shape turns out to be an algebraic curve $C$ is disguise. Namely, the limit shape is essentially the graph of the function

$$\Re \int_{x_0}^{x} dS,$$

where $dS$ is the Seiberg–Witten differential. Thus all ingredients of the answer appear very naturally in the proof.

Random matrix theory and philosophy had many successes in mathematics and physics. Here we have an example when random partitions, while structurally resembling random matrices, offer several advantages. First, the transformation into a random partition problem is geometrically natural and exact. Second, the discretization inherent in partitions regularizes several analytic issues. For further examples along these lines the reader may consult [33].

**1.3. Higher dimensions.** The translation of the gauge theory problem into a random partition problem is explained in Section 2. In Section 3, we analyze the latter problem, in particular, derive and solve the variational problem for the limit shape. Section 4 summarizes parallel results for 3-dimensional partitions, where similar algebraic properties of limit shapes are now proven in great generality.

The surprising fact that free energy $\mathcal{F}$ is given in terms of periods of a hidden algebraic curve $C$ is an example of *mirror symmetry*. A general program of interpreting mirror partners as limit shapes was initiated in [34]. Known results about the limit shapes of periodically weighted 3-dimensional partitions, together with the

conjectural equality of Gromov–Witten and Donaldson–Thomas theories of projective algebraic 3-folds [22] can be interpreted as a verification of this program for toric Calabi–Yau 3-folds. See [35] for an introduction to these ideas.

Note that something completely different is expected to happen in dimensions $> 3$, where the behavior of both random interfaces and Gromov–Witten invariants changes qualitatively.

## 2. The gauge theory problem

**2.1. Instantons.** We begin by recalling some basic facts, see [9] for an excellent mathematical treatment and [8], [10], [44] for a physical one. This will serve as motivation for the introduction of Nekrasov's partitions function in (8) below.

In gauge theories, interactions are transmitted by gauge fields, that is, unitary connections on appropriate vector bundles. In coordinates, these are matrix-valued functions $A_i(x)$ that define covariant derivatives

$$\nabla_i = \frac{\partial}{\partial x_i} + A_i(x), \quad A_i^* = -A_i.$$

We consider the most basic case of the trivial bundle $\mathbb{R}^4 \times \mathbb{C}^r$ over the flat Euclidean space-time $\mathbb{R}^4$, where such coordinate description is global.

The natural (Yang–Mills) energy functional for gauge fields is $L^2$-norm squared $\|F\|^2$ of the curvature

$$F = \sum \left[\nabla_i, \nabla_j\right] dx_i \wedge dx_j.$$

The path integral in quantum gauge theory then takes the form

$$\int_{\text{connections}/\mathcal{G}} \mathcal{D}A \ \exp\left(-\beta \|F\|^2\right) \times \cdots, \tag{2}$$

where dots stand for terms involving other fields of the theory and $\mathcal{G}$ is the group of gauge transformations $g : \mathbb{R}^4 \to U(r)$ acting by

$$\nabla \mapsto g \nabla g^{-1}.$$

In these notes, we will restrict ourselves to pure gauge theory, which is already quite challenging due to the complicated form of the energy. A parallel treatment of certain matter fields can be found in [29].

Our goal is to study (2) as function of the parameter $\beta$ (and boundary conditions at infinity, see below). A head-on probabilistic approach to this problem would be to make it a theory of many interacting random matrices through a discretization of space-time. This is a fascinating topic about which I have nothing to say. In a different direction, when $\beta \gg 0$, the minima of $\|F\|^2$ should dominate the integral.

In *supersymmetric* gauge theory, there is a way to make such approximation exact, thereby reducing the path integral to the following finite-dimensional integrals.

Local minima of $\|F\|^2$ are classified by a topological invariant $c_2 \in \mathbb{Z}$,

$$c_2 = \frac{1}{8\pi^2}\int_{\mathbb{R}^4} \operatorname{tr} F^2,$$

called *charge*, and satisfy a system of first order PDEs

$$F \pm \star F = 0, \tag{3}$$

where $\star$ is the Hodge star operator on 2-forms on $\mathbb{R}^4$. With the plus sign, (3) corresponds to $c_2 > 0$ and is called the anti-self-duality equation. Its solutions are called *instantons*. Minima with $c_2 < 0$ are obtained by reversing the orientation of $\mathbb{R}^4$.

The ASD equations (3) are conformally invariant and can be transported to a punctured 4-sphere $S^4 = \mathbb{R}^4 \cup \{\infty\}$ via stereographic projection. From the removable singularities theorem of Uhlenbeck it follows that any instanton on $\mathbb{R}^4$ extends, after a gauge transformation, to an instanton on $S^4$. Thus we can talk about the value of an instanton at infinity.

Let $\mathcal{G}_0$ be the group of maps $g : S^4 \to U(r)$ such that $g(\infty) = 1$. Modulo $\mathcal{G}_0$, instantons on $S^4$ with $c_2 = n$ are parametrized by a smooth manifold $\mathcal{M}(r, n)$ of real dimension $4rn$. Naively, one would like the contribution from charge $n$ instantons to (2) to be the volume of $\mathcal{M}(r, n)$ in a natural symplectic structure. However, $\mathcal{M}(r, n)$ is noncompact (and its volume is infinite) for two following reasons.

Approximately, an element of $\mathcal{M}(r, n)$ can be imagined as a nonlinear superposition of $n$ instantons of charge 1. Some of those may become point-like, i.e. their curvature may concentrate in a $\delta$-function spike, while others may wander off to infinity. A partial compactification of $\mathcal{M}(r, n)$, constructed by Uhlenbeck, which replaces point-like instanton by just points of $\mathbb{R}^4$, takes care of the first problem but not the second. Nekrasov's idea was to use *equivariant integration* to regularize the instanton contributions.

**2.2. Equivariant regularization.** The group

$$K = \mathrm{SU}(2) \times \mathrm{SU}(r)$$

acts on $\mathcal{M}(r, n)$ by rotations of $\mathbb{R}^4 = \mathbb{C}^2$ and constant gauge transformation, respectively. Our plan is to use this action for regularization. Let us start with the following simplest example: suppose we want to regularize the volume of $\mathbb{R}^2$. A gentle way to do it is to introduce a Gaussian well

$$\int_{\mathbb{R}^2} e^{-t\pi(x^2+y^2)} dx\, dy = \frac{1}{t}, \quad \Re t \geq 0 \tag{4}$$

and thus an effective cut-off at the $|t|^{-1/2}$ scale. Note that the Hamiltonian flow on $\mathbb{R}^2$ generated by $H = \frac{1}{2}(x^2 + y^2)$ with respect to the standard symplectic form

$\omega = dx \wedge dy$ is rotation about the origin with angular velocity one. This makes (4) a simplest instance of the Atiyah–Bott–Duistermaat–Heckman *equivariant localization* formula [2]. We will use localization in the following complex form.

Let $T = \mathbb{C}^*$ act on a complex manifold $X$ with isolated fixed points $X^T$. Suppose that the action of $U(1) \subset T$ is generated by a Hamiltonian $H$ with respect to a symplectic form $\omega$. Then

$$\int_X e^{\omega - 2\pi t H} = \sum_{x \in X^T} \frac{e^{-2\pi t H(x)}}{\det t|_{T_x X}}, \tag{5}$$

where $t$ should be viewed as an element of $\mathrm{Lie}(T) \cong \mathbb{C}$, an so it acts in the complex tangent space $T_x X$ to a fixed point $x \in X$. While (5) is normally stated for compact manifolds $X$, example (4) shows that with care it can work for noncompact ones, too. Scaling both $\omega$ and $H$ to zero, we get from (5) a formal expression

$$\int_X 1 \stackrel{\text{def}}{=} \sum_{x \in X^T} \frac{1}{\det t|_{T_x X}}, \tag{6}$$

which does not depends on the symplectic form and vanishes if $X$ is compact.

A theorem of Donaldson identifies instantons with *holomorphic bundles* on $\mathbb{C}^2 = \mathbb{R}^4$ and thus gives a complex description of $\mathcal{M}(r, n)$. Concretely, $\mathcal{M}(r, n)$ is the moduli space of rank $r$ holomorphic bundles $\mathcal{E} \to \mathbb{CP}^2$ with given 2nd Chern class $c_2(\mathcal{E}) = n$ and a given trivialization along the line

$$L_\infty = \mathbb{CP}^2 \setminus \mathbb{C}^2$$

at infinity. Note that existence of such trivialization implies that $c_1(\mathcal{E}) = 0$. A similar but larger moduli space $\overline{\mathcal{M}}(r, n)$ of *torsion-free sheaves*, see e.g. [12], [24], is a smooth partial compactification of $\mathcal{M}(r, n)$.

The complexification of $K$

$$K_{\mathbb{C}} = \mathrm{SL}(2) \times \mathrm{SL}(r)$$

acts on $\overline{\mathcal{M}}(r, n)$ by operating on $\mathbb{C}^2$ and changing the trivialization at infinity. Equivariant localization with respect to a general $t \in \mathrm{Lie}(K)$

$$t = (\mathrm{diag}(-i\varepsilon, i\varepsilon), \mathrm{diag}(ia_1, \ldots, ia_r)) \tag{7}$$

combines the two following effects. First, it introduces a spatial cut-off parameter $\varepsilon$ as in (4). Second, it introduces dependence on the instanton's behavior at infinity through the parameters $a_i$. While the first factor in $K$ works to shepherd run-away instantons back to the origin, the second works to break the gauge invariance at infinity. In supersymmetric gauge theories, the parameters $a_i$ correspond to the vacuum expectation of the Higgs field and thus are responsible for masses of gauge bosons. In short, they are live physical parameters.

**2.3. Nekrasov partition function.** We are now ready to introduce our main object of study, the partition function of the pure ($\mathcal{N} = 2$ supersymmetric) $U(r)$ gauge theory:

$$Z(\varepsilon; a_1, \dots, a_r; \Lambda) = Z_{\text{pert}} \sum_{n \geq 0} \Lambda^{2rn} \int_{\bar{\mathcal{M}}(r,n)} 1, \tag{8}$$

where the integral is defined by (6) applied to (7),

$$\Lambda = \exp(-4\pi^2 \beta / r),$$

and $Z_{\text{pert}}$ is a certain perturbative factor to be discussed below. The series in (8) is denoted $Z_{\text{inst}}$. Because of factorials in denominators, see (24), $Z_{\text{inst}}$ converges whenever we avoid zero denominators, that is, on the complement of

$$a_i - a_j \equiv 0 \mod \varepsilon. \tag{9}$$

In essence, these factorials are there because the instantons are unordered. Also note that $Z$ is an even function of $\varepsilon$ and a symmetric function of the $a_i$'s.

Since by our regularization rule

$$\operatorname{vol} \mathbb{R}^4 = \int_{\mathbb{R}^4} 1 = \frac{1}{\varepsilon^2},$$

we may expect that as $\varepsilon \to 0$

$$\ln Z(\varepsilon; a; \Lambda) \sim -\frac{1}{\varepsilon^2} \mathcal{F}(a; \Lambda),$$

where $\mathcal{F}$ is the *free energy*. At first, the poles (9) of $Z_{\text{inst}}$, which are getting denser and denser, may look like a problem. Indeed, poles of multiplicity $O(\varepsilon^{-1})$ may affect the free energy, but it is a question of competition with the other terms in $Z_{\text{inst}}$, which the pole-free terms win if $|a_i - a_j| \gg 0$. As a result, either by passing to a subsequence of $\varepsilon$, or by restricting summation in $Z_{\text{inst}}$ to the relevant pole-free terms, we obtain a limit

$$\mathcal{F}_{\text{inst}} = -\lim \varepsilon^2 Z_{\text{inst}},$$

which is analytic and monotone far enough from the walls of the Weyl chambers. Recall that Weyl chambers for $\mathrm{SU}(r)$ are the $r!$ cones obtained from

$$\mathsf{C}_+ = \{a_1 > a_2 > \cdots > a_r,\ \textstyle\sum a_i = 0\}$$

by permuting the coordinates. As $|a_i - a_j|$ get small, poles do complicate the asymptotics. This is the origin of cuts in the analytic function $\mathcal{F}(a)$, $a_i \in \mathbb{C}$.

Nekrasov conjectured in [28] that the free energy $\mathcal{F}$ is the *Seiberg–Witten prepotential*, first obtained in [38], [39] through entirely different considerations. It is defined in terms of a certain family of algebraic curves.

**2.4. Seiberg–Witten geometry.** In the affine space of complex polynomials of the form $P(z) = z^r + O(z^{r-2})$ consider the open set $\boldsymbol{U}$ of polynomials such that

$$P(z) = \pm 2\Lambda^r \tag{10}$$

has $2r$ distinct roots. Over $\boldsymbol{U}$, we have a $g$-dimensional family of complex algebraic curves $C$ of genus $g = r - 1$ defined by

$$\Lambda^r \left( w + \frac{1}{w} \right) = P(z), \quad P \in \boldsymbol{U}. \tag{11}$$

The curve (11) is compactified by adding two points $\partial C = \{w = 0, \infty\}$.

Let $\mathsf{M} \subset \boldsymbol{U}$ be the set of $P(z)$ for which all roots of (10) are real. The corresponding curves $C$ are called *maximal* and play a special role, see e.g. [40]. They arise, for example, as spectral curves of a periodic Toda chain [41]. A maximal curve $C$ has $r$ real ovals, as illustrated in Figure 2. Note that for $z \in \mathbb{R}$, $w$ is either real or lies on the unit circle $|w| = 1$.

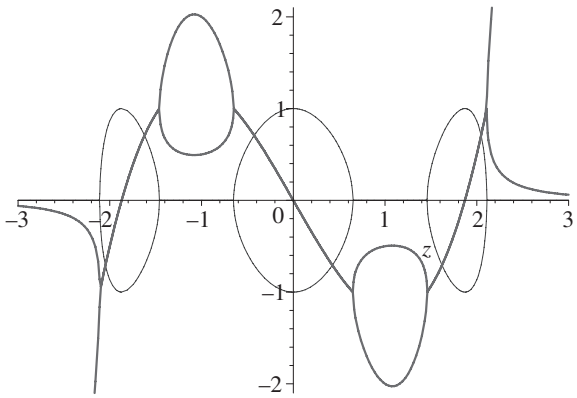



Figure 2. $\Re w$ (bold) and $\Im w$ for $w + 1/w = z^3 - 3.5\,z$ and $z \in \mathbb{R}$.

The intervals $P^{-1}([-2\Lambda^r, 2\Lambda^r]) \subset \mathbb{R}$ on which $|w| = 1$ are called *bands*. The intervals between the bands are called *gaps*. The smaller (in absolute value) root $w$ of the equation (11) can be unambiguously defined for $z \in \mathbb{C} \setminus \{\text{bands}\}$. On the corresponding sheet of the Riemann surface of $w$, we define cycles

$$\alpha_i \in H_1(C - \partial C), \quad \beta_i \in H_1(C, \partial C), \quad i = 1, \dots, r \tag{12}$$

as illustrated in Figure 3, where dotted line means that $\beta_i$ continues on the other sheet. Note that $\alpha_i \cap \beta_j = \delta_{ij}$ and that

$$\overline{\alpha}_i = -\alpha_i, \quad \overline{\beta}_i = \beta_i, \tag{13}$$

where bar stands for complex conjugation. The ovals in Figure 2 represent the cycles $\alpha_i$ and $\beta_i - \beta_{i+1}$.

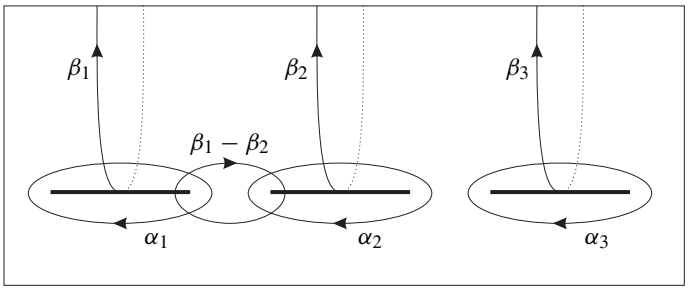



Figure 3. Cycles $\beta_i$ go from $w = \infty$ to $w = 0$. Bold segments indicate bands.

The Seiberg–Witten differential

$$dS = \frac{1}{2\pi i}\, z\, \frac{dw}{w} = \pm\frac{r}{2\pi i}\left(1 + O\left(z^{-2}\right)\right) dz$$

is holomorphic except for a second order pole (without residue) at $\partial C$. Its derivatives with respect to $P \in \boldsymbol{U}$ are, therefore, holomorphic differentials on $C$. In fact, this gives

$$T_P\boldsymbol{U} \cong \text{holomorphic diff. on } C.$$

Nondegeneracy of periods implies the functions

$$a_i \stackrel{\text{def}}{=} \int_{\alpha_i} dS, \quad \sum a_i = 0, \tag{14}$$

which are real on $\mathsf{M}$ by (13), are local coordinates on $\boldsymbol{U}$, as are

$$a_i^\vee - a_{i+1}^\vee \stackrel{\text{def}}{=} 2\pi i \int_{\beta_i - \beta_{i+1}} dS, \quad \sum a_i^\vee = 0. \tag{15}$$

Further, there exists a function $\mathcal{F}(a;\, \Lambda)$, which is real and convex on $\mathsf{M}$, such that

$$\left(\frac{\partial}{\partial a_i} - \frac{\partial}{\partial a_{i+1}}\right)\mathcal{F} = -\left(a_i^\vee - a_{i+1}^\vee\right). \tag{16}$$

Indeed, the Hessian of $\mathcal{F}$ equals $(-2\pi i)$ times the period matrix of $C$, hence symmetric (and positive definite on $\mathsf{M}$). The function $\mathcal{F}$ is called the *Seiberg–Witten prepotential*. Note that $\mathcal{F}$ is multivalued on $\boldsymbol{U}$ and, in fact, its monodromy played a key role in the argument of Seiberg and Witten. By contrast, $\mathsf{M}$ is simply-connected, indeed

$$a^\vee : \mathsf{M} \to \mathsf{C}_+$$

is a diffeomorphism, see e.g. [18] for a more general result. Note that the periods (15) are the areas enclosed by the images of real ovals of $C$ under $(z, w) \mapsto (z, \ln|w|)$. A similar geometric interpretation of the $a_i$'s will be given in (40) below. In particular, the range

$$\mathsf{A} = a(\mathsf{M})$$

of the coordinates (14) is a proper subset of $\mathsf{C}_- = -\mathsf{C}_+$. At infinity of $\mathsf{M}$, we have

$$a_i \sim \{\text{roots of } P\}, \quad a_1 \ll a_2 \ll \cdots \ll a_r.$$

**2.5. Main result.** We have now defined all necessary ingredients to confirm Nekrasov's conjecture in the following strong form:

**Theorem 1** ([29]). *For $a \in \mathsf{A}$,*

$$-\lim_{\varepsilon\to 0} \varepsilon^2 \ln Z(\varepsilon; a; \Lambda) = \mathcal{F}(a; \Lambda), \tag{17}$$

*where $\mathcal{F}$ is the Seiberg–Witten prepotential* (16).

At the boundary of $\mathsf{A}$, free energy has a singularity of the form

$$\mathcal{F} = -\left(a_i^\vee - a_j^\vee\right)^2 \ln\left(a_i^\vee - a_j^\vee\right) + \cdots$$

where dots denote analytic terms. This singularity is one of the main physical features of the Seiberg–Witten theory.

In broad strokes, the logic of the proof was explained in the Introduction. We now proceed with the details.

## 3. The random partition problem

**3.1. Fixed points contributions.** A rank 1 torsion-free sheaf on $\mathbb{C}^2$ is a fancy name to call an ideal $I$ of $\mathbb{C}[x, y]$. Any partition $\lambda$ defines one by

$$I_\lambda = (x^{\lambda_1}, x^{\lambda_2} y, x^{\lambda_3} y^2, \dots) \subset \mathbb{C}[x, y].$$

It is easy to see that all torus-fixed points of $\overline{\mathcal{M}}(r, n)$ have the form

$$\mathfrak{F} = \bigoplus_{k=1}^{r} I_{\lambda^{(k)}}, \quad \sum \left|\lambda^{(k)}\right| = n, \tag{18}$$

where $\lambda^{(k)}$ is an $r$-tuple of partitions. Our goal now is to compute the character of the torus action in the tangent space to the fixed point (18) and thus the contribution of $\mathfrak{F}$ to the sum in (6).

By construction of $\overline{\mathcal{M}}(r, n)$, its tangent space at $\mathfrak{F}$ equals $\operatorname{Ext}^1_{\mathbb{P}^2}(\mathfrak{F}, \mathfrak{F}(-L_\infty))$. From the vanishing of the other Ext-groups we conclude

$$\operatorname{tr} e^t|_{\operatorname{Ext}^1_{\mathbb{P}^2}(\mathfrak{F},\mathfrak{F}(-L_\infty))} = \mathcal{X}_{\mathcal{O}^{\oplus r}}(t) - \mathcal{X}_{\mathfrak{F}}(t), \tag{19}$$

where $\mathcal{X}_{\mathfrak{F}}(t)$ is the character

$$\mathcal{X}_{\mathfrak{F}}(t) = \operatorname{tr} e^t|_{\chi_{\mathbb{C}^2}(\mathfrak{F},\mathfrak{F})}$$

of the infinite-dimensional virtual representation

$$\chi_{\mathbb{C}^2}(\mathfrak{F},\mathfrak{F}) = \operatorname{Ext}^0_{\mathbb{C}^2}(\mathfrak{F},\mathfrak{F}) - \operatorname{Ext}^1_{\mathbb{C}^2}(\mathfrak{F},\mathfrak{F}) + \operatorname{Ext}^2_{\mathbb{C}^2}(\mathfrak{F},\mathfrak{F}).$$

Any graded free resolution of $\mathfrak{F}$ gives

$$\mathcal{X}_{\mathfrak{F}}(t) = |\mathsf{G}_{\mathfrak{F}}(t)|^2, \quad t \in \operatorname{Lie}(K),$$

where $\mathsf{G}_{\mathfrak{F}}(t)$ is, up to a factor, the character of $\mathfrak{F}$ itself

$$\begin{aligned}\mathsf{G}_{\lambda^{(1)},\dots,\lambda^{(r)}}(t) &= (e^{-i\varepsilon/2} - e^{i\varepsilon/2}) \operatorname{tr} e^t|_{\mathfrak{F}} \\ &= \sum_{k=1}^{r} e^{ia_k} \sum_{j=1}^{\infty} \exp\bigl(i\varepsilon\bigl(\lambda_j^{(k)} - j + \tfrac{1}{2}\bigr)\bigr). \end{aligned} \tag{20}$$

It is also a natural generating function of the $r$-tuple $\lambda^{(k)}$.

Note that the weight of any $\mathfrak{F}$ is real and positive, being a product of purely imaginary numbers in conjugate pairs.

**3.2. Perturbative factor.** In the spirit of the original uncompactified gauge theory problem on $\mathbb{R}^4$, we would like to drop the first term in (19) and declare its contribution canceled by $Z_{\text{pert}}$. In view of (20), this requires a regularization of the following product

$$Z_{\text{pert}} \text{“=”} \prod_{k,k'=1}^{r} \prod_{j,j'=1}^{\infty} i(a_k - a_{k'} + \varepsilon(j - j')).$$

A natural regularization is provided by Barnes' double $\Gamma$-function (21), see e.g. [37]. For $c_1, c_2 \in \mathbb{R}$ and $\Re w \gg 0$, define

$$\zeta_2(s; w \mid c_1, c_2) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{dt}{t}\, t^s \, \frac{e^{-wt}}{\prod(1 - e^{-c_i t})}.$$

This has a meromorphic continuation in $s$ with poles at $s = 1, 2$. Define

$$\Gamma_2(w \mid c_1, c_2) = \exp \frac{d}{ds}\, \zeta(s; w \mid c_1, c_2)\big|_{s=0}. \tag{21}$$

Through the difference equation

$$w\, \Gamma_2(w)\, \Gamma_2(w + c_1 + c_2) = \Gamma_2(w + c_1)\, \Gamma_2(w + c_2) \tag{22}$$

it extends to a meromorphic function of $w$. We define

$$Z_{\text{pert}} = \prod_{k,k'} \Gamma_2\left(\frac{i(a_k - a_{k'})}{\Lambda} \,\middle|\, \frac{i\varepsilon}{\Lambda}, \frac{-i\varepsilon}{\Lambda}\right)^{-1}. \tag{23}$$

where $\Gamma_2$ is analytically continued to imaginary arguments using

$$\Gamma_2\,(Mw \mid Mc, -Mc) = M^{\frac{w^2}{2c^2}-\frac{1}{12}}\ \Gamma_2(w \mid c, -c), \quad M \notin (-\infty, 0].$$

The scaling by $\Lambda$ is introduced in (23) to make (8) homogeneous of degree 0 in $a$, $\varepsilon$, and $\Lambda$. Note also

$$\Gamma_2(0 \mid 1, -1) = e^{-\zeta'(-1)}.$$

Our renormalization rule (23) fits nicely with the following transformation of the partition function $Z$.

**3.3. Dual partition function.** For $r = 1$, the weight of $I_\lambda$ in (8) equals

$$\Lambda^{2n}\ \det{}^{-1} t\big|_{T_{I_\lambda}\bar{\mathcal{M}}(1,n)} = \frac{1}{n!}\left(\frac{\Lambda^2}{\varepsilon^2}\right)^n \mathfrak{M}_{\text{Planch}}(\lambda), \tag{24}$$

where $\mathfrak{M}_{\text{Planch}}$ is the Plancherel measure (1) and the prefactor is the Poisson weight with parameter $\Lambda^2/\varepsilon^2$. For $r > 1$, we will transform $Z$ into the partition function (29) of the Plancherel measure in a *periodic potential* with period $r$.

Let a function $\xi : \mathbb{Z} + \frac{1}{2} \to \mathbb{R}$ be periodic with period $r$ and mean 0. The energy $\Xi(\lambda)$ of the configuration $\mathfrak{S}(\lambda)$ in the potential $\xi$ is defined by Abel's rule

$$\Xi(\lambda) = \sum_{x\in\mathfrak{S}(\lambda)} \xi(x) \stackrel{\text{def}}{=} \lim_{z\to+0} \sum_{x\in\mathfrak{S}(\lambda)} \xi(x)\, e^{zx}.$$

Grouping the points of $\mathfrak{S}(\lambda)$ modulo $r$ uniquely determines an $r$-tuple of partitions $\lambda^{(k)}$, known as $r$-quotients of $\lambda$, and shifts $s_k \in \mathbb{Q}$ such that

$$\mathfrak{S}(\lambda) = \bigsqcup_{k=1}^{r} r\big(\mathfrak{S}\big(\lambda^{(k)}\big) + s_k\big) \tag{25}$$

and

$$rs \equiv \rho \mod r\,\mathbb{Z}_0^r, \quad \rho = \left(\tfrac{r-1}{2}, \dots, \tfrac{1-r}{2}\right),$$

where $\mathbb{Z}_0^r$ denotes vectors with zero sum. It follows from (25) that

$$\mathsf{G}_\lambda\,(\varepsilon/r) = \mathsf{G}_{\lambda^{(1)},\dots,\lambda^{(r)}}(\varepsilon;\, \varepsilon s). \tag{26}$$

Letting $\varepsilon \to 2\pi i k$, $k = 1, \dots, r-1$, in (26) gives

$$\Xi(\lambda) = (s, \xi) = \sum s_i\, \xi_i, \quad \xi_i = \xi\left(\tfrac{1}{2} - i\right), \tag{27}$$

while the $\varepsilon \to 0$ limit in (26) yields

$$|\lambda| = r\left(\sum |\lambda^{(k)}| + \sum \frac{s_k^2}{2}\right) + \frac{1-r^2}{24}.$$

Using these formulas and the difference equation (22), we compute

$$Z^\vee(\varepsilon;\xi_1,\dots,\xi_r;\Lambda) \stackrel{\text{def}}{=} \sum_{a\in\varepsilon(\rho+r\mathbb{Z}_0^r)} \exp\left(\frac{(\xi,a)}{r\varepsilon^2}\right) Z(r\varepsilon;a;\Lambda) \tag{28}$$

$$= e^{\zeta'(-1)+\frac{\pi i}{24}} \sum_\lambda \left|\frac{\Lambda}{\varepsilon}\right|^{2|\lambda|-\frac{1}{12}} \left(\frac{\dim\lambda}{|\lambda|!}\right)^2 \exp\left(\frac{\Xi(\lambda)}{\varepsilon}\right). \tag{29}$$

We call (28) the *dual partition function*. By (29), it equals the partition function of a periodically weighted Plancherel measure on partitions.

While it will play no role in what follows, it may be mentioned here that $Z^\vee$ is a very interesting object to study not asymptotically but exactly. For example, Toda equation for $\ln \mathbb{Z}^\vee$ may be found in Section 5 of [29].

**3.4. Dual free energy.** Define the dual free energy by

$$\mathcal{F}^\vee(\xi;\Lambda) = -\lim_{\varepsilon\to 0} \varepsilon^2 \ln Z^\vee. \tag{30}$$

Since (28) is a Riemann sum for Laplace transform, we may expect that

$$\mathcal{F}^\vee(\xi;\Lambda) = \min_{a\in\mathbb{R}_0^r} \frac{1}{r^2}\mathcal{F}(a;\Lambda) - \frac{1}{r}(\xi,a) \tag{31}$$

that is, up to normalization, $\mathcal{F}^\vee$ is the Legendre transform of $\mathcal{F}$. This is because the asymptotics of Laplace transform is determined by one point – the maximum. Our plan is apply to same logic to the infinite-dimensional sum (29), namely, to show that its $\varepsilon \to 0$ asymptotics is determined by a single term, the *limit shape*.

The law of large numbers, a basic principle of probability, implies that on a large scale most random system are deterministic: solids have definite shape, fluids obey the laws of hydrodynamics, etc. Only magnification reveals the full randomness of nature.

In the case at hand, the weight of a partition $\lambda$ in (29), normalized by the whole sum, defines a probability measure on the set of partitions. This measure depends on a parameter $\varepsilon$ and as $\varepsilon \to 0$ it clearly favors partitions of larger and larger size. In fact, the expected size of $\lambda$ grows as $\varepsilon^{-2}$. We thus expect the diagram of $\lambda$, scaled by $\varepsilon$ in both directions, to satisfy a law of large numbers, namely, to have a nonrandom limit shape. By definition, this limit shape will dominate the leading $\varepsilon \to 0$ asymptotics of $Z^\vee$. In absence of the periodic potential $\Xi$, such analysis is a classical result of Logan–Shepp and Vershik–Kerov [21], [42], [43].

Note that the maximum in (31) is over all of $a$, including the problematic region where $|a_i - a_j|$ get small. However, this region does not contribute to $\mathcal{F}^\vee$ as the convexity of free energy is lost there. We will see this reflected in the following properties of $\mathcal{F}^\vee$: it is strictly concave, analytic in the interior of the Weyl chambers, and singular along the chambers' walls.

**3.5. Variational problem for the limit shape.** The *profile* of a partition $\lambda$ is, by definition, the piecewise linear function plotted in bold in Figure 1. Let $\psi_\lambda$ be the profile of $\lambda$ scaled by $\varepsilon$ in both directions. The map $\lambda \mapsto \psi_\lambda$ embeds partitions into the convex set $\Psi$ of functions $\psi$ on $\mathbb{R}$ with Lipschitz constant 1 and

$$|\psi| = \int |\psi(x) - |x||\, dx < \infty.$$

The Lipschitz condition implies

$$\|\psi_1 - \psi_2\|_C \le \|\psi_1 - \psi_2\|_{L^1}^{1/2}, \quad \psi_1, \psi_2 \in \Psi,$$

and so $\Psi$ is complete and separable in the $L^1$-metric. Some function of a partition have a natural continuous extension to $\Psi$, for example

$$\mathsf{G}_\lambda(\varepsilon) = \frac{1}{e^{i\varepsilon/2} - e^{-i\varepsilon/2}} \left(1 - \frac{1}{2}\int e^{ix}\,(\psi_\lambda(x) - |x|)\, dx\right),$$

while others, specifically the ones appearing in (29), do not. An adequate language for dealing with this is the following.

Let $f(\lambda) \ge 0$ be a function on partitions depending on the parameter $\varepsilon$. We say that it satisfies a *large deviation* principle with action (rate) functional $\mathcal{S}_f(\psi)$ if for any set $A \subset \Psi$

$$-\lim \varepsilon^2 \ln \sum_{\psi_\lambda \in A} f(\lambda) \subset \left[\inf_{\bar{A}} \mathcal{S}_f, \inf_{A^\circ} \mathcal{S}_f\right] \subset \mathbb{R} \cup \{+\infty\}, \tag{32}$$

where lim denotes all limit points, $A^\circ$ and $\bar{A}$ stand for the interior and closure of $A$, respectively.

For the Plancherel weight (24), Logan–Shepp and Vershik–Kerov proved a large deviation principle with action

$$\mathcal{S}_{\mathrm{pl}}(\psi) = \frac{1}{2}\int_{x<y} (1 + \psi'(x))(1 - \psi'(y)) \ln \frac{|x-y|}{\Lambda}\, dx\, dy. \tag{33}$$

Note that in this case the sum in (32) may be replaced by maximum because the number of partitions of $n$ grows subexponentially in $n$. In other words, there is no entropic contribution in (33).

The periodic potential $\Xi(\lambda)$ produces a *surface tension* addition to the total action $\mathcal{S}$

$$\mathcal{S} = \mathcal{S}_{\mathrm{pl}} + \mathcal{S}_{\mathrm{surf}}, \quad \mathcal{S}_{\mathrm{surf}}(\psi) = \frac{1}{2}\int \sigma(\psi')\, dt,$$

where $\sigma$ is a convex piecewise-linear function of the kind plotted in Figure 4. It is linear on segments of length $2/r$ with slopes $\{\xi_i\}$, in increasing order. The form of $\mathcal{S}_{\mathrm{surf}}$ is easy to deduce directly; it can also be seen as e.g. the most degenerate case of

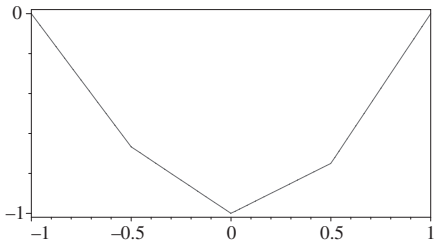



Figure 4. The surface tension $\sigma$ for $r = 4$ and $\xi = \{-\frac{4}{3}, -\frac{2}{3}, \frac{1}{2}, \frac{3}{2}\}$.

the surface tension formula from Section 4.2. The singularities in the surface tension $\sigma$ are responsible for *facets*, that is, linear pieces, in the minimizer, see Figure 6. The slopes of these facets are precisely the points where $\sigma'$ is discontinuous[1]. Note that $\sigma$ and hence $\mathcal{I}$ is a symmetric function of the $\xi_i$'s.

The functional $\mathcal{I}$ is strictly convex and its sublevel sets $\mathcal{I}^{-1}((-\infty, c])$ are compact, which can be seen by rewriting $\mathcal{I}_{\mathrm{pl}}$ in terms of the Sobolev $H^{1/2}$ norm, see [21], [43]. Therefore it has a unique minimum $\psi_\star$ – the limit shape. The large deviation principle and the definition of the dual free energy and (30) together imply

$$\mathcal{F}^{\vee}(\xi; \Lambda) = \mathcal{I}(\psi_\star). \tag{34}$$

Our business, therefore, is to find this minimizer $\psi_\star$.

**3.6. The minimizer.** By convexity, a local minimum of $\mathcal{I}$ is automatically a global one. Since $\sigma$ has one-sided derivatives, a local minimum can be characterized by nonnegativity of all directional derivatives. This leads to the following *complementary slackness* conditions for the convolution of $\psi_\star''(x)$ with the kernel

$$\mathsf{L}(x) = x \ln \frac{|x|}{\Lambda} - x = \int_0^x \ln \left| \frac{y}{\Lambda} \right| dy.$$

There exists a constant $c_0$, which is the Lagrange multiplier from the constraint $\int \delta\psi' = 0$, such that

$$\begin{aligned} &\mathsf{L} * \psi_\star''(x) + c_0 = \xi_i, && \psi_\star'(x) \in \left(-1 + \tfrac{2i-2}{r}, -1 + \tfrac{2i}{r}\right), \\ &\mathsf{L} * \psi_\star''(x) + c_0 \in [\xi_i, \xi_{i+1}], && \psi_\star'(x) = -1 + \tfrac{2i}{r}, \end{aligned} \tag{35}$$

where to simplify notation we assumed that

$$\xi \in \mathsf{C}_-, \quad \xi_0 = -\infty, \quad \xi_{r+1} = +\infty.$$

[1]There are many advantages in viewing random partitions as 2-dimensional slices of random 3-dimensional objects discussed in Section 4. From the probability viewpoint, this links random partitions with rather realistic models of crystalline surfaces with local interaction, enriching both techniques and intuition. In particular, coexistence of facets and curved regions in our limit shapes is the same phenomenon as observed in natural crystals. From the gauge theory viewpoint, it is also very natural, especially in the context of 5-dimensional theory on $\mathbb{R}^4 \times S^1$, which corresponds to the $K$-theory of the instanton moduli spaces.

Recall that $\mathsf{C}_-$ denotes the negative Weyl chamber.

The function $\psi_\star''$ will turn out to be nonnegative and supported on a union of $r$ intervals, which are precisely the *bands* of Section 2.4. The gaps will produce the facets in the limit shape.

It is elementary to see that for a maximal curve (11) the map

$$\Phi(z) = 1 + \frac{2}{\pi i r} \ln w = 1 + \frac{2}{\pi i} \ln \frac{\Lambda}{z} + O(z^{-1}), \quad z \to \infty \tag{36}$$

where $w$ is the smaller root of (11), defines a conformal map of the upper half-plane to a slit half-strip

$$\Delta \subset \{z \,|\Im z > 0, |\Re z| < 1\}$$

as in Figure 5. The slits in $\Delta$ go along

$$\Re z = -1 + 2i/r, \quad i = 1, \dots, r-1,$$

and their lengths are, essentially, the critical values of the polynomial $P(z)$. The bands and gaps are preimages of the horizontal and vertical segments of $\partial\Delta$, respectively.

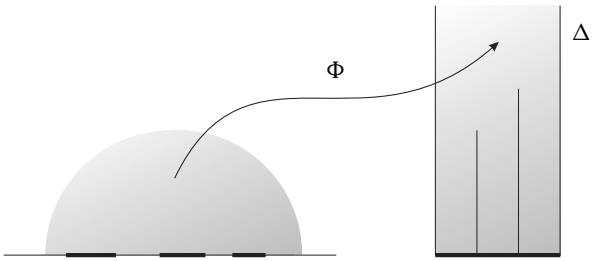



Figure 5. Conformal map defined by a maximal curve.

We claim that

$$\psi_\star' = \Re\Phi\big|_{\mathbb{R}}, \tag{37}$$

where the polynomial $P(z)$ is determined by the relation (39) below. The equations (35) are verified for (37) as follows. Since $\Phi'(z) = O(z^{-1})$, $z \to \infty$, we have the Hilbert transform relation

$$\text{P.V.}\,\frac{1}{x} * \Re\Phi'\big|_{\mathbb{R}} = \pi\Im\Phi'\big|_{\mathbb{R}}.$$

Integrating it once and using (36) to fix the integration constant, we get

$$(\mathsf{L} * \Re\Phi')' = \pi\Im\Phi.$$

Therefore, the function $\mathsf{L} * \Re\Phi'$ is constant on the bands and strictly increasing on the gaps, hence (37) satisfies (35) with

$$\xi_{i+1} - \xi_i = \pi \int_{i\text{th gap}} \Im\Phi(x)\,dx \tag{38}$$

Integrating (38) by parts and using definitions from Section 2.4 gives

$$\xi = -\frac{a^{\vee}}{r}, \tag{39}$$

thus every limit shape $\psi_\star$ comes from a maximal curve. For example, the limit shape corresponding to the curve from Figure 2 is plotted in Figure 6. Note also that for $C \in \mathsf{M}$, we have

$$a_i = \frac{r}{2}\left(I_{i-1} - I_i\right), \tag{40}$$

where $I_i$ is the intercept of the $i$th facet of the limit shape. In particular, $\mathsf{A} \subset \mathsf{C}_-$.

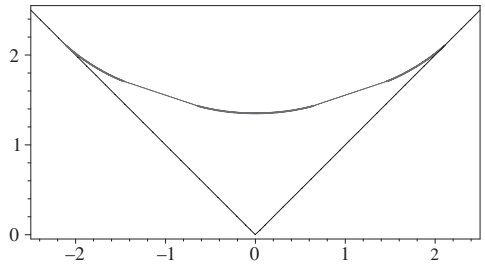



Figure 6. Limit shape corresponding to the curve from Figure 2. Thin segments are facets.

For given $\xi \in \mathsf{C}_-$, consider the distribution of the $r$-quotients $\lambda^{(i)}$ of the partition $\lambda$, as defined in Section 3.3. For the shifts $s_k$ in (25) we have using (27)

$$\varepsilon s \to -\frac{\partial \mathcal{F}^{\vee}}{\partial \xi}, \quad \varepsilon \to 0,$$

in probability. Observe that

$$\frac{\partial}{\partial \xi}\mathcal{F}^{\vee}(\xi) = \left[\frac{\partial}{\partial \xi}\mathcal{S}\right](\psi_\star)$$

since the other term, containing $\frac{\partial}{\partial \xi}\psi_\star$, vanishes by the definition of a maximum. Definitions and integration by parts yield

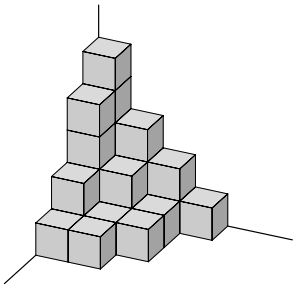
$$-\left(\frac{\partial}{\partial \xi_i} - \frac{\partial}{\partial \xi_{i+1}}\right)\mathcal{F}^{\vee} = \frac{a_i - a_{i+1}}{r}.$$

By (26), this means that the resulting sum over the $r$-quotients $\lambda^{(i)}$ is the original partition function $Z$ with parameters $a \in \mathsf{A}$. This concludes the proof.

## 4. The next dimension

### 4.1. Stepped surfaces

An obvious 3-dimensional generalization of a partition, also known as a plane partition can be seen on the right. More generally, we consider *stepped surfaces*, that is, continuous surfaces glued out of sides of a unit cube, spanning a given polygonal contour in $\mathbb{R}^3$, and projecting 1-to-1 in the (1, 1, 1) direction, see Figure 7. Note that stepped surfaces minimize the surface area for given boundary conditions, hence can be viewed as zero temperature limit of the interface in the 3D Ising model.



The most natural measure on stepped surfaces is the uniform one with given boundary conditions, possibly conditioned on the volume enclosed. It induces Plancherel-like measures on 2-dimensional slices. Stepped surfaces are in a natural bijection with fully packed dimers on the hexagonal lattice and Kasteleyn theory of planar dimers [16] forms the basis of most subsequent developments.

The following law of large numbers for stepped surfaces was proven in [7]. Let $C_n$ be a sequence of boundary contours such that each $C_n$ can be spanned by at least one stepped surface. Suppose that $n^{-1}C_n$ converge to a given curve $C \subset \mathbb{R}^3$. Then, scaled by $n^{-1}$, uniform measures on stepped surfaces spanning $C_n$ converge to the $\delta$-measure on a single Lipschitz surface spanning $C$ – the limit shape. This limit shape formation is clearly visible in Figure 7.

Figure 7. A limit shape simulation. The frozen boundary is the inscribed cardioid.

The limit shape is the unique minimizer of the following functional. Let the surface be parameterized by $x_3 = h(x_3 - x_1, x_3 - x_2)$, where $h$ is a Lipschitz function

with gradient in the triangle $\triangle$ with vertices $(0, 0)$, $(0, 1)$, $(1, 0)$. Let $\Omega$ be the planar region enclosed by the projection of $C$ in the $(1, 1, 1)$ direction. We will use $(x, y) = (x_3 - x_1, x_2 - x_1)$ as coordinates on $\Omega$. The limit shape is the unique minimizer of

$$\mathcal{S}_{\text{step}}(h) = \int_\Omega \sigma_{\text{step}}(\nabla h)\, dx\, dy, \tag{41}$$

where, in the language of [20], the surface tension $\sigma_{\text{step}}$ is the Legendre dual of the Ronkin function of the straight line

$$z + w = 1. \tag{42}$$

We recall that for a plane curve $P(z, w) = 0$, its Ronkin function [23] is defined by

$$R(x, y) = \frac{1}{(2\pi i)^2} \iint_{\substack{|z|=e^x \\ |w|=e^y}} \log|P(z, w)|\, \frac{dz}{z}\, \frac{dw}{w}. \tag{43}$$

The gradient $\nabla R$ always takes values in the Newton polygon $\Delta(P)$ of the polynomial $P$, so $\Delta(P)$ is naturally the domain of the Legendre transform $R^\vee$. For the straight line as above, the Newton polygon is evidently the triangle $\triangle$.

The surface tension $\sigma_{\text{step}}$ is singular and not strictly convex at the boundary of $\triangle$, which leads to formation of *facets* and *edges* in the limit shape (which can be clearly seen in Figure 7). This models facet formation in natural interfaces, e.g. crystalline surfaces, and is the most interesting aspect of the model. Note that facets are completely ordered (or *frozen*). The boundary between the ordered and disordered (or *liquid*) regions is known as the *frozen boundary*.

The following transformation of the Euler-Lagrange equation for (41) found in [19] greatly facilitates the study of the facet formation. Namely, in the liquid region we have

$$\nabla h = \frac{1}{\pi}(\arg w, -\arg z), \tag{44}$$

where the functions $z$ and $w$ solve the differential equation

$$\frac{z_x}{z} + \frac{w_y}{w} = c \tag{45}$$

and the algebraic equation (42). Here $c$ is the Lagrange multiplier for the volume constraint $\int_\Omega h = \text{const}$, the unconstrained case is $c = 0$. At the boundary of the liquid region, $z$ and $w$ become real and the $\nabla h$ starts to point in one of the coordinate directions.

The first-order quasilinear equation (45) is, essentially, the complex Burgers equation $z_x = z z_y$ and, in particular, it can be solved by complex characteristics as follows. There exists an analytic function $Q(z, w)$ such that

$$Q(e^{-cx} z, e^{-cy} w) = 0. \tag{46}$$

In other words, $z(x, y)$ can be found by solving (42) and (46). In spirit, this is very close to Weierstraß parametrization of minimal surfaces in terms of analytic data.

Frozen boundary can only develop if $Q$ is real, in which case the roots $(z, w)$ and $(\bar{z}, \bar{w})$ of (46) coincide at the frozen boundary. At a smooth point of the frozen boundary, the multiplicity of this root will be exactly two, hence $\nabla h$ has a square-root singularity there. As a result, the limit shape has an $x^{3/2}$ singularity at the generic point of the frozen boundary, thus recovering the well-known Pokrovsky–Talapov law [36] in this situation. At special points of the frozen boundary, triple solutions of (46) occur, leading to a cusp singularity. One such point can be seen in Figure 7.

Remarkably, for a dense set of boundary condition the function $Q$ is, in fact, a polynomial. Consequently, the frozen boundary takes the form $R(e^{cx}, e^{cy}) = 0$, where $R$ is the polynomial defining the planar dual of the curve $Q = 0$. This allows to use powerful tools of algebraic geometry to study the singularities of the solutions, see [19]. The precise result proven there is

**Theorem 2** ([19]). *Suppose the boundary contour $C$ is a connected polygon with $3k$ sides in coordinate directions* (*cyclically repeated*) *which can be spanned by a Lipschitz function with gradient in $\triangle$. Then $Q = 0$ is an algebraic curve of degree $k$ and genus zero.*

For example, for the boundary contour in Figure 7 we have $k = 3$ (one of the boundary edges there has zero length) and hence $R$ is the dual of a degree 3 genus 0 curve – a cardioid. The procedure of determining $Q$ from the boundary conditions is effective and can be turned into a practical numeric homotopy procedure, see [19]. Higher genus frozen boundaries occur for multiply-connected domains, in fact, the genus of $Q$ equals the genus of the liquid region.

Of course, for a probabilist, the law of large numbers is only the beginning and the questions about CLT corrections to the limit shape and local statistics of the surface in various regions of the limit shape follow immediately. Conjecturally, the limit shape controls the answers to all these questions. For example, the function $e^{-cx}z$ defines a *complex structure* on the liquid region and, conjecturally, the Gaussian correction to the limit shape is given by the *massless free field* in the corresponding conformal structure. In the absence of frozen boundaries and without the volume constraint, this is proven in [17]. See e.g. [15], [17], [20], [32] for an introduction to the local statistics questions.

**4.2. Periodic weights.** Having discussed periodically weighted Plancherel measure and a 3-dimensional analog of the Plancherel measure, we now turn to periodically weighted stepped surfaces. This is very natural if stepped surfaces are interpreted as crystalline interfaces. Periodic weights are introduced as follows: we weight each square by a periodic function of $x_3 - x_1$ and $x_2 - x_1$ (with some integer period $M$).

The role previously played by the straight line (42) is now played by a certain higher degree curve $P(z, w) = 0$, the spectral curve of the corresponding periodic

Kasteleyn operator. In particular, the surface tension $\sigma_{\text{step}}$ is now replaced by the Legendre dual of the Ronkin function of $P$, see [20]. We have

$$\deg P = M$$

and the coefficients of $P$ depend polynomially on the weights.

The main result of [20], known as *maximality*, says that for real and positive weights the curve $P$ is always a real algebraic curve of a very special kind, namely, a *Harnack curve*, see [23]. Conversely, as shown in [18], all Harnack curves arise in this way.

Harnack curves are, in some sense, the best possible real curves; their many remarkable properties are discussed in [23]. One of several equivalent definitions of a Harnack curve is that the map

$$(z, w) \mapsto (\log|z|, \log|w|) \tag{47}$$

from $P(z, w) = 0$ to $\mathbb{R}^2$ is 1-to-1 on the real locus of $P$ and 2-to-1 over the rest. The image of $P = 0$ under (47) is known as the *amoeba* of $P$. Note from (43) that the gradient $\nabla R$ of the Ronkin function of $P$ is nonconstant precisely for $(x, y)$ in the amoeba of $P$. In other words, the Ronkin function has a facet (that is, a linear piece) over every component of the amoeba complement. The 2-to-1 property implies that the number of compact facets of Ronkin function equals the (geometric) genus of the curve $P$. Each of these facets translates into the singularity of the surface tension and, hence, into facets with the same slope in limit shapes.

By Wulff's theorem, the Ronkin function itself is a minimizer, corresponding to its own ("crystal corner") boundary conditions. An example of the Ronkin function of a genus 1 Harnack curve can be seen in Figure 8.

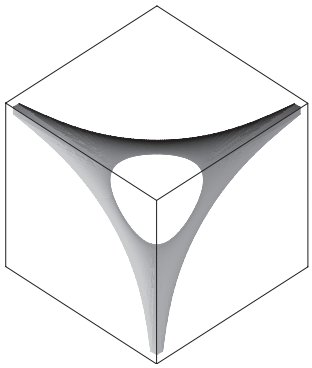


Figure 8. The (curved part of the) Ronkin function of a genus 1 curve. Its projection to the plane is the amoeba.

Maximality implies *persistence of facets*, namely, for fixed period $M$, there will be $\binom{M-1}{2}$ compact facets of the Ronkin function and $\binom{M-1}{2}$ corresponding singularities

of the surface tension, except on a codimension 2 subvariety of the space of weights. It also implies e.g. the following *universality of height fluctuations* in the liquid region

$$\operatorname{Var}(h(a) - h(b)) \sim \frac{1}{\pi} \ln \|a - b\|, \quad \|a - b\| \to \infty.$$

Remarkably, formulas (44), (45), and (46) need no modifications for periodic weights. Replacing (42) by $P(z, w) = 0$ is the only change required, see [19].

From our experience with periodically weighted Plancherel measure, it is natural to expect that, for some special boundary conditions, the partition function of periodically weighted stepped surfaces will encode valuable physical information. A natural choice of "special boundary conditions" are the those of a *crystal corner*, when we require the surface to be asymptotic to given planes at infinity, as in Figure 8. For convergence of the partition function, one introduces a fugacity factor $q^{\mathrm{vol}}$, where the missing volume is measured with respect to the "full corner".

I hope that further study will reveal many special properties of such crystal corner partition functions. Their extremely degenerate limits have been identified with all-genera, all-degree generating functions for Donaldson–Thomas invariants of toric Calabi–Yau threefolds. Namely, as the periodic weights become extreme, all limit shapes, and the Ronkin function in particular, degenerate to piecewise linear functions. This is known as the *tropical limit*. The only remaining features of limit shapes are the edges and the triple points, where 2 and 3 facets meet, respectively. In this tropical limit, the partition function becomes the partition function of ordinary, unweighted, 3D partitions located at triple points. These 3D partitions may have infinite legs along the edges, as in Figure 9 and through these legs they interact with their neighbors. This description precisely matches the localization formula for Donaldson–Thomas invariants of the toric threefold whose toric polyhedron is given by the piecewise linear limit shape, see [22].

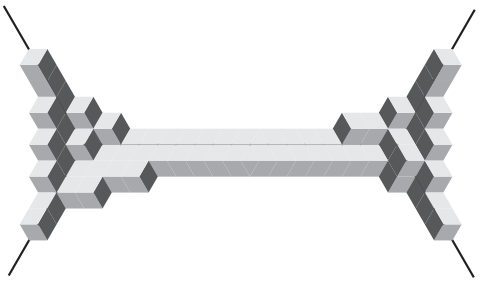


Figure 9. Two 3D partitions connected at an angle through an infinite leg.

Donaldson–Thomas theory of any 3-fold has been conjectured to be equivalent, in a nontrivial way, to the Gromov–Witten theory of the same 3-fold in [22]. For the toric Calabi–Yau 3-folds, this specializes to the earlier *topological vertex* conjecture of [1]. It is impossible to adequately review this subject here, see [35] for an introduction.

This is also related to the supersymmetric gauge theories considered in Section 2, or rather their 5-dimensional generalizations, via a procedure called *geometric engineering* of gauge theories. See for example [13] and references therein.

I find such close and unexpected interaction between rather basic statistical models and instantons in supersymmetric gauge and string theories very exciting and promising. The field is still full of wide open questions and, in my opinion, it is also full of new phenomena waiting to be discovered.

## References


[1] Aganagic, M., Klemm, A., Marino, M., Vafa, C., The Topological Vertex. *Comm. Math. Phys.* **254** (2005), 425–478.

[2] Atyiah, M., Bott, R., The moment map and equivariant cohomology. *Topology* **23** (1) (1984), 1–28.

[3] Baik, J., Deift, P., Johansson, K., On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12** (1999), 1119–1178.

[4] Borodin, A., Okounkov, A., Olshanski, G., Asymptotics of the Plancherel measures for symmetric groups. *J. Amer. Math. Soc.* **13** (3) (2000), 481–515.

[5] Braverman, A., Instanton counting via affine Lie algebras. I. Equivariant $J$-functions of (affine) flag manifolds and Whittaker vectors. In *Algebraic structures and moduli spaces*, CRM Proc. Lecture Notes 38, Amer. Math. Soc., Providence, RI, 2004, 113–132.

[6] Braverman, A., Etingof, P., Instanton counting via affine Lie algebras II: from Whittaker vectors to the Seiberg-Witten prepotential. math.AG/0409441.

[7] Cohn, H., Kenyon, R., Propp, J., A variational principle for domino tilings. *J. Amer. Math. Soc.* **14** (2) (2001), 297–346.

[8] D'Hoker, E., Phong, D., Lectures on supersymmetric Yang-Mills theory and integrable systems. In *Theoretical physics at the end of the twentieth century*, CRM Ser. Math. Phys., Springer-Verlag, New York 2002, 1–125.

[9] Donaldson, S, Kronheimer, P., *The geometry of four-manifolds*. Oxford Math. Monogr., The Clarendon Press, New York 1990.

[10] Dorey, N., Hollowood, T., Khoze, V., Mattis, M., The calculus of many instantons. *Phys. Rep.* **371** (4–5) (2002), 231–459.

[11] Hollowood, T., Iqbal, A., Vafa, C., Matrix models, geometric engineering, and elliptic genera. hep-th/0310272.

[12] Huybrechts, D., Lehn, M., *The geometry of moduli spaces of sheaves*. Aspects Math. E31, Vieweg, Braunschweig 1997.

[13] Iqbal,A., Kashani-Poor, A.-K., The vertex on a strip. hep-th/0410174.

[14] Johansson, K., Discrete orthogonal polynomial ensembles and the Plancherel measure. *Ann. of Math.* **153** (1) (2001), 259–296.

[15] Johansson, K., Random matrices and determinantal processes. math-ph/0510038.

[16] Kasteleyn, P., Graph theory and crystal physics. In *Graph Theory and Theoretical Physics*, Academic Press, London 1967, 43–110.

[17] Kenyon, R., Height fluctuations in honeycomb dimers. math-ph/0405052.

[18] Kenyon, R., Okounkov, A., Planar dimers and Harnack curves. math.AG/0311062.

[19] Kenyon, R., Okounkov, A., Limit shapes and complex Burgers equation. math-ph/0507007.

[20] Kenyon, R., Okounkov, A., Sheffield, S., Dimers and amoebae. math-ph/0311005.

[21] Logan, B., Shepp, L., A variational problem for random Young tableaux. *Adv. Math.* **26** (1977), 206–222.

[22] Maulik, D., Nekrasov, N., Okounkov, A., Pandharipande, R., Gromov-Witten theory and Donaldson-Thomas theory, I. & II. math.AG/0312059, math.AG/0406092.

[23] Mikhalkin, G., Amoebas of algebraic varieties and tropical geometry. In *Different faces of geometry*, Int. Math. Ser. (N. Y.), Kluwer/Plenum, New York 2004, 257–300.

[24] Nakajima, H., *Lectures on Hilbert schemes of points on surfaces*. Univ. Lecture Ser. 18, Amer. Math. Soc., Providence, RI, 1999.

[25] Nakajima, H., Yoshioka, K., Instanton counting on blowup. I. 4-dimensional pure gauge theory. *Invent. Math.* **162** (2005), 313–355.

[26] Nakajima, H., Yoshioka, K., Lectures on instanton counting. *Algebraic structures and moduli spaces*, CRM Proc. Lecture Notes, 38, Amer. Math. Soc., Providence, RI, 2004, 31–101.

[27] Nakajima, H., Yoshioka, K., Instanton counting on blowup. II. $K$-theoretic partition function. math.AG/0505553.

[28] Nekrasov, N., Seiberg-Witten prepotential from instanton counting. *Adv. Theor. Math. Phys.* **7** (5) (2003), 831–864.

[29] Nekrasov, N., Okounkov, A., Seiberg-Witten Theory and Random Partitions. In *The Unity of Mathematics* (ed. by P. Etingof, V. Retakh, I. M. Singer), Progr. Math. 244, Birkhäuser, Boston, MA, 2006, 525–596.

[30] Nekrasov, N., Shadchin, S., ABCD of instantons. *Comm. Math. Phys.* **252** (2004), 359–391.

[31] Okounkov, A., Random matrices and random permutations. *Internat. Math. Res. Notices* **2000** (20) (2000), 1043–1095.

[32] Okounkov, A., Symmetric function and random partitions. In *Symmetric functions 2001: surveys of developments and perspectives* (ed. by S. Fomin), Kluwer Acad. Publ., Dordrecht 2002, 223–252.

[33] Okounkov, A., The uses of random partitions. math-ph/0309015.

[34] Okounkov, A., Reshetikhin, N., Vafa, C., Quantum Calabi-Yau and Classical Crystals. In *The Unity of Mathematics* (ed. by P. Etingof, V. Retakh, I. M. Singer), Progr. Math. 244, Birkhäuser, Boston, MA, 2006, 597–618.

[35] Okounkov, A., Random surfaces enumerating algebraic curves. In *Proceedings of Fourth European Congress of Mathematics*, EMS, Zürich 2005, 751–768.

[36] Pokrovsky, V., Talapov, A., Theory of two-dimensional incommensurate crystals. *Soviet Phys. JETP* **78** (1) (1980), 269–295.

[37] Ruijsenaars, S., On Barnes' multiple zeta and gamma functions. *Adv. Math.* **156** (1) (2000), 107–132.

[38] Seiberg, N., Witten, E., Electric-magnetic duality, monopole condensation, and confinement in $\mathcal{N} = 2$ supersymmetric Yang-Mills theory. *Nuclear Phys. B* **426** (1994), 19–52; Erratum ibid. **430** (1994), 485–486.

[39] Seiberg, N., Witten, E., Monopoles, duality and chiral symmetry breaking in $\mathcal{N} = 2$ supersymmetric QCD. *Nuclear Phys. B* **431** (1994), 484–550.

[40] Sodin, M., Yuditskii, P., Functions that deviate least from zero on closed subsets of the real axis. *St. Petersburg Math. J.* **4** (2) (1993), 201–249.

[41] Toda, M., *Theory of nonlinear lattices*. Springer Ser. Solid-State Sci. 20, Springer-Verlag, Berlin 1981.

[42] Vershik, A., Kerov, S., Asymptotics of the Plancherel measure of the symmetric group and the limit form of Young tableaux. *Soviet Math. Dokl.* **18** (1977), 527–531.

[43] Vershik, A., Kerov, S., Asymptotics of maximal and typical dimensions of irreducible representations of a symmetric group. *Funct. Anal. Appl.* **19** (1) (1985), 21–31 .

[44] Witten, E., Dynamics of quantum field theory. In *Quantum fields and strings: a course for mathematicians* (ed. by P. Deligne, P. Etingof, D. Freed, L. Jeffrey, D. Kazhdan, J. Morgan, D. Morrison and E. Witten), Vol. 2, Amer. Math. Soc., Providence, RI, IAS, Princeton, NJ, 1999, 1119–1424.

Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, New Jersey, 08544, U.S.A.

E-mail: okounkov@math.princeton.edu

# Estimation in inverse problems and second-generation wavelets

Dominique Picard and Gérard Kerkyacharian

**Abstract.** We consider the problem of recovering a function $f$ when we receive a blurred (by a linear operator) and noisy version: $Y_\varepsilon = Kf + \varepsilon \dot{W}$. We will have as guides 2 famous examples of such inverse problems: the deconvolution and the Wicksell problem. The direct problem ($K$ is the identity) isolates the denoising operation. It cannot be solved unless accepting to estimate a smoothed version of $f$: for instance, if $f$ has an expansion on a basis, this smoothing might correspond to stopping the expansion at some stage $m$. Then a crucial problem lies in finding an equilibrium for $m$, considering the fact that for $m$ large, the difference between $f$ and its smoothed version is small, whereas the random effect introduces an error which is increasing with $m$. In the true inverse problem, in addition to denoising, we have to 'inverse the operator' $K$, an operation which not only creates the usual difficulties, but also introduces the necessity to control the additional instability due to the inversion of the random noise. Our purpose here is to emphasize the fact that in such a problem there generally exists a basis which is fully adapted to the problem, where for instance the inversion remains very stable: this is the singular value decomposition basis. On the other hand, the SVD basis might be difficult to determine and to numerically manipulate. It also might not be appropriate for the accurate description of the solution with a small number of parameters. Moreover, in many practical situations the signal provides inhomogeneous regularity, and its local features are especially interesting to recover. In such cases, other bases (in particular, localised bases such as wavelet bases) may be much more appropriate to give a good representation of the object at hand. Our approach here will be to produce estimation procedures keeping the advantages of a localisation properly without loosing the stability and computability of SVD decompositions. We will especially consider two cases. In the first one (which is the case of the deconvolution example) we show that a fairly simple algorithm (WAVE-VD), using an appropriate thresholding technique performed on a standard wavelet system, enables us to estimate the object with rates which are almost optimal up to logarithmic factors for any $\mathbb{L}_p$ loss function and on the whole range of Besov spaces. In the second case (which is the case of the Wicksell example where the SVD basis lies in the range of Jacobi polynomials) we prove that a similar algorithm (NEED-VD) can be performed provided one replaces the standard wavelet system by a second generation wavelet-type basis: the needlets. We use here the construction (essentially following the work of Petrushev and co-authors) of a localised frame linked with a prescribed basis (here Jacobi polynomials) using a Littlewood–Paley decomposition combined with a cubature formula. Section 5 describes the direct case ($K = I$). It has its own interest and will act as a guide for understanding the 'true' inverse models for a reader who is not familiar with nonparametric statistical estimation. It can be read first. Section 1 introduces the general inverse problem and describes the examples of deconvolution and Wicksell's problem. A review of standard methods is given with a special focus on SVD methods. Section 2 describes the WAVE-VD procedure. Section 3 and 4 give a description of the needlets constructions and the performances of the NEED-VD procedure.

**Mathematics Subject Classification (2000).** 62G07, 62G20, 62C20.

**Keywords.** Nonparametric estimation, denoising, inverse models, thresholding, Meyer wavelet, singular value decomposition, Littlewood–Paley decomposition.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

## 1. Inverse models

Let $\mathbb{H}$ and $\mathbb{K}$ be two Hilbert spaces. $K$ is a linear operator: $f \in \mathbb{H} \mapsto Kf \in \mathbb{K}$. The standard linear ill-posed inverse problem consists of recovering a good approximation $f_\varepsilon$ of $f$, solution of

$$g = Kf, \tag{1}$$

when only a perturbation $g_\varepsilon$ of $g$ is observed. In this paper we will consider the case where this perturbation is an additive stochastic white noise. Namely, we observe $Y_\varepsilon$ defined by the following equation:

$$Y_\varepsilon = Kf + \varepsilon \dot{W}, \ \mathbb{H}, \ \mathbb{K}, \tag{2}$$

where $\varepsilon$ is the amplitude of the noise. It is supposed to be a small parameter which will tend to 0. Our error will be measured in terms of this small parameter.

$\dot{W}$ is a $\mathbb{K}$-white noise: i.e. for any $g, h$ in $\mathbb{K}$, $\xi(g) := (\dot{W}, g)_{\mathbb{K}}$, $\xi(h) := (\dot{W}, h)_{\mathbb{K}}$ form a random gaussian vector, centered, with marginal variance $\|g\|^2_{\mathbb{K}}$, $\|h\|^2_{\mathbb{K}}$, and covariance $(g, h)_{\mathbb{K}}$ (with the obvious extension when one considers $k$ functions instead of 2).

Equation (2) means that for any $g$ in $\mathbb{K}$, we observe $Y_\varepsilon(g) := (Y_\varepsilon, g)_{\mathbb{K}} = (Kf, g)_{\mathbb{K}} + \varepsilon\xi(g)$ where $\xi(g) \sim N(0, \|g\|^2)$, and $Y_\varepsilon(g)$, $Y_\varepsilon(h)$ are independent random variables for orthogonal functions $g$ and $h$.

The case where $K$ is the identity is called the 'direct model' and is summarized as a memento in Section 5. The reader who is unfamiliar with nonparametric statistical estimation is invited to consult this section, which will act as a guide for understanding the more general inverse models. In particular it is recalled therein that the model (2) is in fact an approximation of models appearing in real practical situations, for instance the case where (2) is replaced by a discretisation.

### 1.1. Two examples: the problem of deconvolution and Wicksell's problem

**1.1.1. Deconvolution.** The following problem is probably one of the most famous among inverse problems in signal processing. In the deconvolution problem we consider the following operator. In this case let $\mathbb{H} = \mathbb{K}$ be the set of square integrable periodic functions with the standard $\mathbb{L}_2([0, 1])$ norm and consider

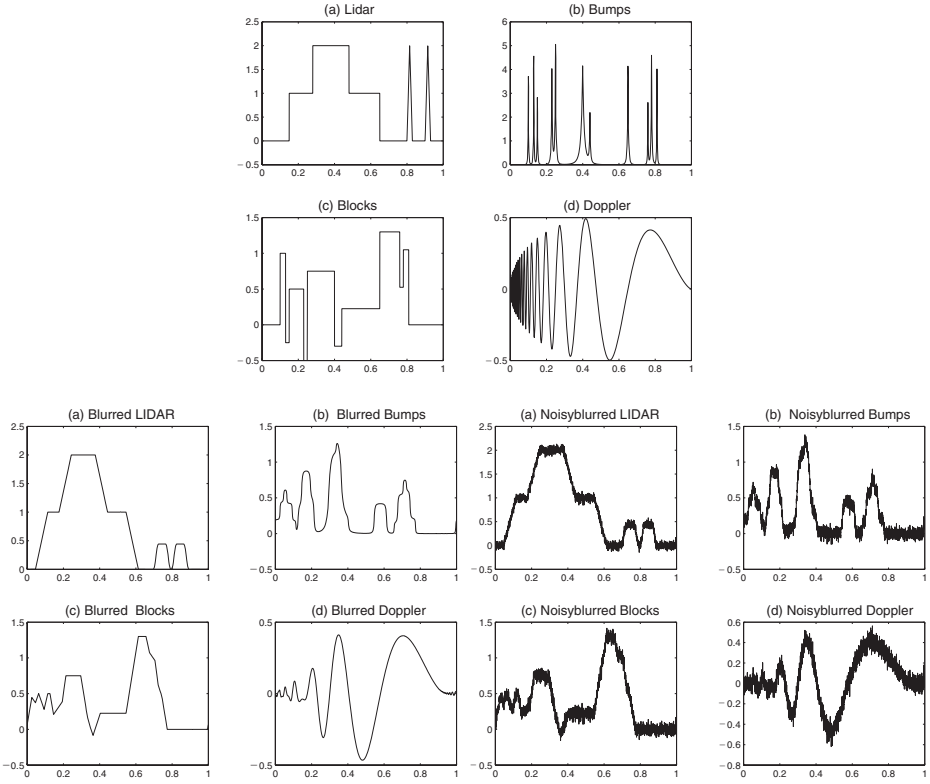
$$f \in \mathbb{H} \mapsto Kf = \int_0^1 \gamma(u - t) f(t)\, dt \in \mathbb{H}, \tag{3}$$

where $\gamma$ is a known function of $\mathbb{H}$, which is generally assumed to be a regular function (often in the sense that its Fourier coefficients $\hat{\gamma}_k$ behave like $k^{-\nu}$). A very common example is also the box-car function: $\gamma(t) = \frac{1}{2a}\mathbb{I}\{[-a, a]\}(k)$.

The following figures show first four original signals to recover, which are well-known test-signals of the statistical literature. They provide typical features which are difficult to restore: bumps, blocks and Doppler effects. The second and third series of

pictures show their deformation after blurring (i.e. convolution with a regular function) and addition of a noise. These figures show how the convolution regularizes the signal, making it very difficult to recover, especially the high frequency features. A statistical investigation of these signals can be found in [22].

A variant of this problem consists in observing $Y_1, \ldots, Y_n$, $n$ independent and identically distributed random variables where each $Y_i$ may be written as $Y_i = X_i + U_i$, where $X_i$ and $U_i$ again are independent, the distribution of $U_i$ is known and of density $\gamma$ and we want to recover the common density of the $X_i$'s. The direct problem is the case where $U_i = 0$, for all $i$, and is corresponding to a standard density estimation problem (see Section 5.1) . Hence the variables $U_i$ are acting as perturbations of the $X_i$'s, whose density is to be recovered.




**1.1.2. Wicksell's problem.** Another typical example is the following classical Wicksell problem [42]. Suppose a population of spheres is embedded in a medium. The spheres have radii that may be assumed to be drawn independently from a density $f$. A random plane slice is taken through the medium and those spheres that are intersected by the plane furnish circles the radii of which are the points of observation $Y_1, \ldots, Y_n$. The unfolding problem is then to infer the density of the sphere radii from the observed circle radii. This unfolding problem also arises in medicine, where

the spheres might be tumors in an animal's liver [36], as well as in numerous other contexts (biological, engineering,…), see for instance [9].

Following [42] and [23], Wicksell's problem corresponds to the following operator:

$$\mathbb{H} = \mathbb{L}_2([0,1], d\mu)\ d\mu(x) = (4x)^{-1}dx,$$
$$\mathbb{K} = \mathbb{L}_2([0,1], d\lambda)\ d\lambda(x) = 4\pi^{-1}(1-y^2)^{1/2}dy$$
$$Kf(y) = \frac{\pi}{4}y(1-y^2)^{-1/2}\int_y^1 (x^2-y^2)^{-1/2}f(x)d\mu.$$

Notice, however, that in this presentation, again in order to avoid additional technicalities, we handle this problem in the white noise framework, which is simpler than the original problem expressed above in density terms.

**1.2. Singular value decomposition and projection methods.** Let us begin with a quick description of well-known methods in inverse problems with random noise.

Under the assumption that $K$ is compact, there exist 2 orthonormal bases (SVD bases) $(e_k)$ of $\mathbb{H}$ and $(g_k)$ of $\mathbb{K}$, respectively, and a sequence $(b_k)$, tending to 0 when $k$ goes to infinity, such that

$$Ke_k = b_k g_k, \quad K^* g_k = b_k e_k$$

if $K^*$ is the adjoint operator.

For the sake of simplicity we suppose in the sequel that $K$ and $K^*$ are into. Otherwise we have to take care of the kernels of these operators. The bases $(e_k)$ and $(g_k)$ are called singular value bases, whereas the $b_k$'s are simply called singular values.

**Deconvolution.** In this standard case simple calculations prove that the SVD bases $(e_k)$ and $(g_k)$ both coincide with the Fourier basis. The singular values are corresponding to the Fourier coefficients of the function $\gamma$:

$$b_k = \hat{\gamma}_k. \tag{4}$$

**Wicksell.** In this case, following [23], we have the following SVD:

$$e_k(x) = 4(k+1)^{1/2}x^2 P_k^{0,1}(2x^2-1),$$
$$g_k(y) = U_{2k+1}(y).$$

$P_k^{0,1}$ is the Jacobi polynomial of type (0, 1) with degree $k$, and $U_k$ is the second type Chebyshev polynomial with degree $k$. The singular values are

$$b_k = \frac{\pi}{16}(1+k)^{-1/2}. \tag{5}$$

**1.2.1. SVD method.** The singular value decomposition (SVD) of $K$,

$$Kf = \sum_k b_k \langle f, e_k \rangle g_k,$$

gives rise to approximations of the type

$$f_\varepsilon = \sum_{k=0}^{N} b_k^{-1} \langle y_\varepsilon, g_k \rangle e_k,$$

where $N = N(\varepsilon)$ has to be chosen properly. This SVD method is very attractive theoretically and can be shown to be asymptotically optimal in many situations (see Mathé and Pereverzev [31], Cavalier and Tsybakov [6], Mair and Ruymgaart [29]). It also has the big advantage of performing a quick and stable inversion of the operator. However, it suffers from different types of limitations. The SVD bases might be difficult to determine as well as to numerically manipulate. Secondly, while these bases are fully adapted to describe the operator $K$, they might not be appropriate for the accurate description of the solution with a small number of parameters.

Also in many practical situations the signal provides inhomogeneous regularity, and its local features are especially interesting to recover. In such cases other bases (in particular localised bases such as wavelet bases) may be much more appropriate to give a good representation of the object at hand.

**1.2.2. Projection methods.** Projection methods which are defined as solutions of (1) restricted to finite dimensional subspaces $\mathbb{H}_N$ and $\mathbb{K}_N$ (of dimension $N$) also give rise to attractive approximations of $f$, by properly choosing the subspaces and the tuning parameter $N$ (Dicken and Maass [10], Mathé and Pereverzev [31] together with their non linear counterparts Cavalier and Tsybakov [6], Cavalier et al. [7], Tsybakov [41], Goldenshluger and Pereverzev [19], Efromovich and Koltchinskii [16]). In the case where $\mathbb{H} = \mathbb{K}$ and $K$ is a self-adjoint operator, the system is particularly simple to solve since the restricted operator $K_N$ is symmetric positive definite. This is the so-called Galerkin method. Obviously, restricting to finite subspaces has similar effects and can also be seen as a *Tychonov regularisation*, i.e. minimizing the least square functional penalised by a regularisation term.

The advantage of the Galerkin method is to allow the choice of the basis. However the Galerkin method suffers from the drawback of being unstable in many cases.

Comparing the SVD and Galerkin methods exactly states one main difficulty of the problem. The possible antagonism between the SVD basis where the inversion of the system is easy, and a 'localised' basis where the signal is sparsely represented, will be the issue we are trying to address here.

**1.3. Cut-off, linear methods, thresholding.** *The reader may profitably look at Subsections* 5.3 *and* 5.4*, where the linear methods and thresholding techniques are presented in detail in the direct case.*

SVD as well as Galerkin methods are very sensitive with respect to the choice of the tuning parameter $N(\varepsilon)$. This problem can be solved theoretically. However the solution heavily depends on prior assumptions of regularity on the solution, which have to be known in advance.

In the last ten years, many nonlinear methods have been developed especially in the direct case with the objective of automatically adapting to the unknown smoothness and local singular behavior of the solution. In the direct case, one of the most attractive methods is probably wavelet thresholding, since it allies numerical simplicity to asymptotic optimality on a large variety of functional classes such as Besov or Sobolev classes.

To adapt this approach in inverse problems, Donoho [11] introduced a wavelet-like decomposition, specifically adapted to the operator $K$ (wavelet–vaguelette-decomposition) and provided a thresholding algorithm on this decomposition. In Abramovitch and Silverman [1], this method was compared with the similar vaguelette–wavelet-decomposition. Other wavelet approaches, might be mentioned such as Antoniadis and Bigot [2], Antoniadis et al. [3] and, especially for the deconvolution problem, Penski and Vidakovic [37], Fan and Koo [17], Kalifa and Mallat [24], Neelamani et al. [34].

Later, Cohen et al. [8] introduced an algorithm combining a Galerkin inversion with a thresholding algorithm.

The approach developed in the sequel is greatly influenced by these previous works. The accent we put here is on constructing (when necessary) new generation wavelet-type bases well adapted to the operator $K$, instead of sticking to the standard wavelet bases and reducing the range of potential operators covered by the method.

## 2. Wave-VD-type estimation

We explain here the basic idea of the method, which is very simple. Let us expand $f$ using a well-suited basis ('the wavelet-type' basis', to be defined later):

$$f = \sum (f, \psi_\lambda)_{\mathbb{H}} \psi_\lambda.$$

Using Parseval's identity we have $\beta_\lambda = (f, \psi_\lambda)_{\mathbb{H}} = \sum f_i \psi_\lambda^i$ for $f_i = (f, e_i)_{\mathbb{H}}$ and $\psi_\lambda^i = (\psi_\lambda, e_i)_{\mathbb{H}}$. Let us put $Y_i = (Y_\varepsilon, g_i)_{\mathbb{K}}$. We then have

$$Y_i = (Kf, g_i)_{\mathbb{K}} + \varepsilon\xi_i = (f, K^* g_i)_{\mathbb{K}} + \varepsilon\xi_i = \Big( \sum_j f_j e_j, K^* g_i \Big)_{\mathbb{H}} + \varepsilon\xi_i = b_i f_i + \varepsilon\xi_i,$$

where the $\xi_i$'s are forming a sequence of independent centered gaussian variables with variance 1. Furthermore,

$$\hat{\beta}_\lambda = \sum_i \frac{Y_i}{b_i} \psi_\lambda^i$$

is such that $\mathbb{E}(\hat{\beta}_\lambda) = \beta_\lambda$ (i.e. its average value is $\beta_\lambda$). It is a plausible estimate of $\beta_\lambda$. Let us now put ourselves in a multiresolution setting, taking $\lambda = (j, k)$ for $j \geq 0$, $k$ belonging to a set $\chi_j$, and consider

$$\hat{f} = \sum_{j=-1}^{J} \sum_{k \in \chi_j} t(\hat{\beta}_{jk}) \psi_{jk},$$

where $t$ is a thresholding operator. (*The reader who is unfamiliar with thresholding techniques is referred to Section* 5.4.)

$$t(\hat{\beta}_{jk}) = \hat{\beta}_{jk} I\{|\hat{\beta}_{jk}| \geq \kappa t_\varepsilon \sigma_j\}, \quad t_\varepsilon = \varepsilon \sqrt{\log 1/\varepsilon}, \tag{6}$$

where $I\{A\}$ denotes the indicator function of the set $A^*$. Here $\kappa$ is a tuning parameter of the method which will be properly chosen later. A main difference here with the direct case is the fact that the thresholding is depending on the resolution level through the constant $\sigma_j$ which also will be stated more precisely later. Our main discussion will concern the choice of the basis $(\psi_{jk})$. In particular, we shall see that coherence properties with the SVD basis are of special interest.

We will particularly focus on two situations (corresponding to the two examples discussed in the introduction). In the first type of cases, the operator has as SVD bases the Fourier basis. In this case, this 'coherence' is easily obtained with 'standard' wavelets (still, not any kind of standard wavelet as will be seen). However, more difficult problems (and typically Wicksell's problem) require, when we need to mix these coherence conditions with the desired property of localisation of the basis, the construction of new objects: second generation-type wavelets.

**2.1. WAVE-VD in a wavelet scenario.** In this section we take $\{\psi_{jk},\ j \geq -1,\ k \in \chi_j\}$ to be a standard wavelet basis. More precisely, we suppose as usual that $\psi_{-1}$ stands for the scaling function and, for any $j \geq -1$, $\chi_j$ is a set of order $2^j$ contained in $\mathbb{N}$. Moreover, we assume that the following properties are true. There exist constants $c_p, C_p, d_p$ such that

$$c_p 2^{j(\frac{p}{2}-1)} \leq \|\psi_{jk}\|_p^p \leq C_p 2^{j(\frac{p}{2}-1)}, \tag{7}$$

$$\Big\| \sum_{k \in \chi_j} u_k \psi_{jk} \Big\|_p^p \leq D_p \sum_{k \in \chi_j} |u_k^p| \|\psi_{jk}\|_p^p \quad \text{for any sequence } u_k. \tag{8}$$

It is well known (see for instance Meyer [32]) that wavelet bases provide characterisations of smoothness spaces such as Hölder spaces $\mathrm{Lip}(s)$, Sobolev spaces $W_p^s$ as well as Besov spaces $B_{pq}^s$ for a range of indices $s$ depending on the wavelet $\psi$. For the scale of Besov spaces which includes as particular cases $\mathrm{Lip}(s) = B_{\infty\infty}^s$ (if $s \notin \mathbb{N}$) and $W_p^s = B_{pp}^s$ (if $p = 2$), the characterisation has the following form:

$$\text{If } f = \sum_{j \geq -1} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}, \text{ then } \|f\|_{B_{pq}^s} \sim \Big\| \big(2^{j[s+\frac{1}{2}-\frac{1}{p}]} \|\beta_{j\cdot}\|_{l_p}\big)_{j \geq -1} \Big\|_{l_q}. \tag{9}$$

As in Section 5, we consider the loss of a decision $\hat{f}$ if the truth is $f$ as the $\mathbb{L}_p$ norm $\|\hat{f} - f\|_p$, and its associated risk

$$\mathbb{E}\|\hat{f} - f\|_p^p.$$

Here $\mathbb{E}$ denotes the expectation with respect to the random part of the observation $y_\varepsilon$. The following theorem is going to evaluate this risk, when the strategy is the one introduced in the previous section, and when the true function belongs to a Besov ball ($f \in B^s_{\pi,r}(M) \iff \|f\|_{B^s_{pq}} \le M$). One nice property of this estimation procedure is that it does not need the a priori knowledge of this regularity to get a good rate of convergence. If $(e_k)$ is the SVD basis introduced in Section 1.2, $b_k$ are the singular values and $\psi^i_{jk} = \langle e_i, \psi_{jk}\rangle$, we consider the estimator $\hat{f}$ defined in the beginning of Section 2.

**Theorem 2.1.** *Assume that* $1 < p < \infty$, $2\nu + 1 > 0$ *and*

$$\sigma_j^2 := \sum_i \left[\frac{\psi^i_{jk}}{b_i}\right]^2 \le C 2^{2j\nu} \quad \textit{for all } j \ge 0. \tag{10}$$

*Put* $\kappa^2 \ge 16p$, $2^J = [t_\varepsilon]^{\frac{-2}{2\nu+1}}$. *If* $f$ *belongs to* $B^s_{\pi,r}(M)$ *with* $\pi \ge 1$, $s \ge 1/\pi$, $r \ge 1$ (*with the restriction* $r \le \pi$ *if* $s = (2\nu+1)\left(\frac{p}{2\pi} - \frac{1}{2}\right)$), *then we have*

$$\mathbb{E}\|\hat{f} - f\|_p^p \le C \log(1/\varepsilon)^{p-1} [\varepsilon^2 \log(1/\varepsilon)]^{\alpha p}, \tag{11}$$

*with*

$$\alpha = \frac{s}{1 + 2(\nu + s)} \qquad \textit{if } s \ge (2\nu+1)\left(\tfrac{p}{2\pi} - \tfrac{1}{2}\right),$$
$$\alpha = \frac{s - 1/\pi + 1/p}{1 + 2(\nu + s - 1/\pi)} \quad \textit{if } \tfrac{1}{\pi} \le s < (2\nu+1)\left(\tfrac{p}{2\pi} - \tfrac{1}{2}\right).$$

**Remarks.** 1. Condition (10) is essential here. As will be shown later, this condition is linking the wavelet system with the singular value decomposition of the kernel $K$. If we set ourselves in the deconvolution case, the SVD basis is the *Fourier* basis in such a way that $\psi^i_{jk}$ is simply the Fourier coefficient of $\psi_{jk}$. If we choose as wavelet basis the periodized Meyer wavelet basis (see Meyer [32] and Mallat [30]), conditions (7) and (8) are satisfied. In addition, as the Meyer wavelet has the remarkable property of being compactly supported in the Fourier domain, simple calculations prove that, for any $j \ge 0$, $k$, the number of $i$'s such that $\psi^i_{jk} \ne 0$ is finite and equal to $2^j$. Then if we assume to be in the so-called 'regular' case ($b_k \sim k^{-\nu}$, for all $k$), it is easy to establish that (10) is true. This condition is also true for more general cases in the deconvolution setting such as the box-car deconvolution, see [22], [27].

2. These results are minimax (see [43]) up to logarithmic factors. This means that if we consider the best estimator in its worst performance over a given Besov

ball, this estimator attains a rate of convergence which is the one given in (11) up to logarithmic factors.

3. If we compare these results with the rates of convergence obtained in the direct model (see Subsections 5.3 and 5.4), we see that the difference (up to logarithmic terms) essentially lies in the parameter $\nu$ which acts as a reducing factor of the rate of convergence. This parameter quantifies the extra difficulty offered by the inverse problem. It is often called coefficient of illposedness. If we recall that in the deconvolution case, the coefficients $b_k$ are the Fourier coefficients of the function $\gamma$, the illposedness coefficient then clearly appears to be closely related to the regularity of the blurring function.

This result has been proved in the deconvolution case in [22]. The proof of the theorem is given in Appendix I.

**2.2. WAVE-VD in Jacobi scenario: NEED-VD.** We have seen that the results given above are true under the condition (10) on the wavelet basis.

Let us first appreciate how the condition (10) links the 'wavelet-type' basis to the SVD basis $(e_k)$. To see this let us put ourselves in the regular case:

$$b_i \sim i^{-\nu}.$$

(By this we mean more precisely that there exist two positive constants $c$ and $c'$ such that $c' i^{-\nu} \leq b_i \leq c i^{-\nu}$.)

If (10) is true, we have

$$C 2^{2j\nu} \geq \sum_m \sum_{2^m \leq i < 2^{m+1}} \left[ \frac{\psi_{jk}^i}{b_i} \right]^2.$$

Hence, for all $m \geq j$,

$$\sum_{2^m \leq i < 2^{m+1}} [\psi_{jk}^i]^2 \leq c 2^{2\nu(j-m)}.$$

This suggests the necessity to construct a 'wavelet-type' basis having support, at the level $j$, with respect to the SVD basis (sum in $i$) concentrated on the integers between $2^j$ and $2^{j+1}$ and exponentially decreasing after this band. This is exactly the case of Meyer's wavelet, when the SVD basis is the Fourier basis.

In the general case of an arbitrary linear operator giving rise to an arbitrary SVD basis $(e_k)$, and if in addition to (10) we add a localisation condition on the basis, we do not know if such a construction can be performed. However, in some cases, even quite as far from the deconvolution as the Wicksell problem, one can build a 'second generation wavelet-type' basis, with exactly these properties.

The following construction due to Petrushev and collaborators ([33], [39], [38]) exactly realizes the paradigm mentioned above, producing a frame (the needlet basis) in the case of Jacobi polynomials (as well as in different other cases such as spherical harmonics, Hermite functions, Laguerre polynomials) which has the property of being localised.

## 3. Petrushev construction of needlets

Frames were introduced in the 1950s by Duffin and Schaeffer [15] to represent functions via over-complete sets. Frames including tight frames arise naturally in wavelet analysis on $\mathbb{R}^d$. Tight frames which are very close to orthonormal bases are especially useful in signal and image processing.

We will see that the following construction has the advantage of being easily computable and producing well-localised tight frames constructed on a specified orthonormal basis.

We recall the following definition.

**Definition 3.1.** Let $\mathbb{H}$ be a Hilbert space. A sequence $(e_n)$ in $\mathbb{H}$ is said to be a *tight frame* (with constant 1) if

$$\|f\|^2 = \sum_n |\langle f, e_n\rangle|^2 \quad \text{for all } f \in \mathbb{H}.$$

Let now $\mathcal{Y}$ be a metric space, $\mu$ a finite measure. Let us suppose that we have the decomposition

$$\mathbb{L}_2(\mathcal{Y}, \mu) = \bigoplus_{k=0}^{\infty} H_k,$$

where the $H_k$'s are finite dimensional spaces. For the sake of simplicity we suppose that $H_0$ is reduced to the constants.

Let $L_k$ be the orthogonal projection on $H_k$:

$$L_k(f)(x) = \int_{\mathcal{Y}} f(y) L_k(x, y) d\mu(y) \quad \text{for all } f \in \mathbb{L}_2(\mathcal{Y}, \mu),$$

where

$$L_k(x, y) = \sum_{i=1}^{l_k} e_i^k(x) \bar{e}_i^k(y),$$

$l_k$ is the dimension of $H_k$ and $(e_i^k)_{i=1,\dots,l_k}$ is an orthonormal basis of $H_k$. Observe that we have the following property of the projection operators:

$$\int L_k(x, y) L_m(y, z) d\mu(z) = \delta_{k,m} L_k(x, z). \tag{12}$$

The construction, also inspired by the paper of Frazier, Jawerth and Weiss [18], is based on two fundamental steps: Littlewood–Paley decomposition and discretization, which are summarized in the following two subsections.

**3.1. Littlewood–Paley decomposition.** Let $\varphi$ be a $C^\infty$ function supported in $|\xi| \le 1$ such that $1 \ge \varphi(\xi) \ge 0$ and $\varphi(\xi) = 1$ if $|\xi| \le \frac{1}{2}$. We define

$$a^2(\xi) = \varphi(\xi/2) - \varphi(\xi) \ge 0$$

so that

$$\sum_j a^2(\xi/2^j) = 1 \quad \text{for all } |\xi| \ge 1. \tag{13}$$

We further define the operator

$$\Lambda_j = \sum_{k\ge 0} a^2(k/2^j) L_k$$

and the associated kernel

$$\Lambda_j(x, y) = \sum_{k\ge 0} a^2(k/2^j) L_k(x, y) = \sum_{2^{j-1}<k<2^{j+1}} a^2(k/2^j) L_k(x, y).$$

The following assertion is true.

**Proposition 3.2.** *For all* $f \in \mathbb{H}$

$$f = \lim_{J\to\infty} L_0(f) + \sum_{j=0}^{J} \Lambda_j(f) \tag{14}$$

*and*

$$\Lambda_j(x, y) = \int M_j(x, z) M_j(z, y) d\mu(z) \quad \textit{for } M_j(x, y) = \sum_k a(k/2^j) L_k(x, y). \tag{15}$$

*Proof.*

$$L_0(f) + \sum_{j=0}^{J} \Lambda_j(f) = L_0 + \sum_{j=0}^{J} \Big( \sum_k a^2(k/2^j) L_k \Big) = \sum_k \varphi(k/2^{J+1}) L_k \tag{16}$$

Hence

$$\begin{aligned} &\Big\| \sum_k \varphi(k/2^{J+1}) L_k(f) - f \Big\|^2 \\ &\quad = \sum_{l\ge 2^{J+1}} \|L_l(f)\|^2 + \sum_{2^J \le l < 2^{J+1}} \|L_l(f)(1 - \varphi(l/2^{J+1})\|^2 \\ &\quad \le \sum_{l\ge 2^J} \|L_l(f)\|^2 \longrightarrow 0, \quad \text{when } J \to \infty. \end{aligned}$$

(15) is a simple consequence of (12). □

**3.2. Discretization.** Let us define

$$\mathcal{K}_k = \bigoplus_{m=0}^{k} H_m,$$

and let us assume that some additional assumptions are true:

1. $f \in \mathcal{K}_k \Longrightarrow \bar{f} \in \mathcal{K}_k$.
2. $f \in \mathcal{K}_k, g \in \mathcal{K}_l \Longrightarrow fg \in \mathcal{K}_{k+l}$.
3. Quadrature formula: for all $k \in \mathbb{N}$, there exists a finite subset $\chi_k$ of $\mathcal{Y}$ and positive real numbers $\lambda_\xi > 0$ indexed by the elements $\xi$ of $\chi_k$ such that

$$\int f\,d\mu = \sum_{\xi\in\chi_k} \lambda_\xi f(\xi) \quad \text{for all } f \in \mathcal{K}_k.$$

Then the operator $M_j$ defined in the subsection above is such that $M_j(x,z) = \overline{M_j(z,x)}$ and

$$z \mapsto M_j(x,z) \in \mathcal{K}_{2^{j+1}-1}.$$

Hence

$$z \mapsto M_j(x,z)M_j(z,y) \in \mathcal{K}_{2^{j+2}-2},$$

and we can write

$$\Lambda_j(x,y) = \int M_j(x,z)M_j(z,y)\,d\mu(z) = \sum_{\xi\in\chi_{2^{j+2}-2}} \lambda_\xi M_j(x,\xi)M_j(\xi,y).$$

This implies

$$\begin{aligned}\Lambda_j f(x) &= \int \Lambda_j(x,y)f(y)\,d\mu(y) = \int \sum_{\xi\in\chi_{2^{j+2}-2}} \lambda_\xi M_j(x,\xi)M_j(\xi,y)f(y)\,d\mu(y)\\ &= \sum_{\xi\in\chi_{2^{j+2}-2}} \sqrt{\lambda_\xi}M_j(x,\xi)\int \overline{\sqrt{\lambda_\xi}M_j(y,\xi)}f(y)\,d\mu(y).\end{aligned}$$

This can be summarized in the following way if we put $\sqrt{\lambda_\xi}M_j(x,\xi) = \psi_{j,\xi}(x)$ and $\chi_{2^{j+2}-2} = \mathbb{Z}_j$:

$$\Lambda_j f(x) = \sum_{\xi\in\mathbb{Z}_j} \langle f, \psi_{j,\xi}\rangle \psi_{j,\xi}(x).$$

**Proposition 3.3.** *The family $(\psi_{j,\xi})_{j\in\mathbb{N},\xi\in\mathbb{Z}_j}$ is a tight frame.*

*Proof.* As

$$f = \lim_{J\to\infty} (L_0(f) + \sum_{j\le J} \Lambda_j(f)),$$

we have

$$\|f\|^2 = \lim_{J\to\infty} (\langle L_0(f), f\rangle + \sum_{j\le J} \langle \Lambda_j(f), f\rangle),$$

but

$$\langle \Lambda_j(f), f\rangle = \sum_{\xi\in\mathbb{Z}_j} \langle f, \psi_{j,\xi}\rangle\langle \psi_{j,\xi}, f\rangle = \sum_{\xi\in\mathbb{Z}_j} |\langle f, \psi_{j,\xi}\rangle|^2,$$

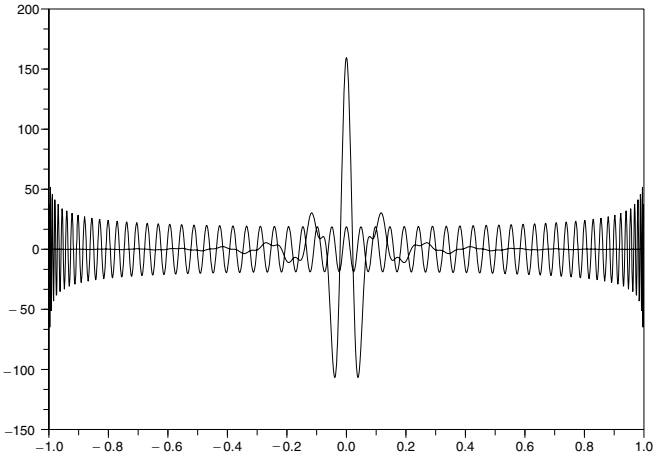
and if $\psi_0$ is a normalized constant we have $\langle L_0(f), f\rangle = |\langle f, \psi_0\rangle|^2$ so that

$$\|f\|^2 = |\langle f, \psi_0\rangle|^2 + \sum_{j\in\mathbb{N},\ \xi\in\mathbb{Z}_j} |\langle f, \psi_{j,\xi}\rangle|^2.$$

But this is exactly the characterization of a tight frame. □

**3.3. Localisation properties.** This construction has been performed in different frameworks by Petrushev and coauthors giving in each situation very nice localisation properties.

The following figure (thanks to Paolo Baldi) is an illustration of this phenomenon: it shows a needlet constructed as explained above using Legendre polynomials of degree $2^8$. The highly oscillating function is a Legendre polynomial of degree $2^8$, whereas the localised one is a needlet centered approximately in the middle of the interval. Its localisation properties are remarkable considering the fact that both functions are polynomials of the same order.



In the case of the sphere of $\mathbb{R}^{d+1}$, where the spaces $H_k$ are spanned by spherical harmonics, the following localisation property is proved in Narcowich, Petrushev and Ward [33]: for any $k$ there exists a constant $C_k$ such that

$$|\psi_{j\eta}(\xi)| \leq \frac{C_k 2^{dj/2}}{[1+2^j \arccos < \eta, \xi >]^k}.$$

A similar result exists in the case of Laguerre polynomials on $\mathbb{R}_+$ [25].

In the case of Jacobi polynomials on the interval with Jacobi weight, the following localisation property is shown by Petrushev and Xu [38]. For any $k$ there exist constants $C$, $c$ such that

$$|\psi_{j\eta}(\cos\theta)| \leq \frac{C 2^{j/2}}{(1+(2^j|\theta - \arccos\eta|)^k \sqrt{w_{\alpha\beta}(2^j, \cos\theta)}},$$

where $w_{\alpha\beta}(n,x) = (1-x+n^{-2})^{\alpha+1/2}(1+x+n^{-2})^{\beta+1/2}$, $-1 \leq x \leq 1$ if $\alpha > -1/2$, $\beta > -1/2$.

## 4. NEED-VD in the Jacobi case

Let us now come back to the estimation algorithm.

We consider the inverse problem (2) with

$$\mathbb{H} = \mathbb{L}_2(I, d\gamma(x)), \quad I = [-1,1], \quad d\gamma(x) = \omega_{\alpha,\beta}(x)dx,$$
$$\omega_{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta; \quad \alpha > -1/2,\ \beta > -1/2.$$

For the sake of simplicity, let us suppose $\alpha \geq \beta$. (Otherwise we can exchange the parameters.)

Let $P_k$ be the normalized Jacobi polynomial for this weight. We suppose that these polynomials appear as SVD basis of the operator $K$, as it is the case for the Wicksell problem with $\beta = 0$, $\alpha = 1$, $b_k \sim k^{-1/2}$.

**4.1. Needlets and condition (10).** Let us define the 'needlets' as constructed above:

$$\psi_{j,\eta_k}(x) = \sum_l \hat{a}(l/2^{j-1}) P_l(x) P_l(\eta_k) \sqrt{b_{j,\eta_k}}. \tag{17}$$

The following proposition asserts that such a construction always implies the condition (10) in the regular case.

**Proposition 4.1.** *Assume that $\psi_{j,\eta_k}$ is a frame. If $b_i \sim i^{-\nu}$ then*

$$\sigma_j^2 := \sum_i \left[\frac{\psi_{jk}^i}{b_i}\right]^2 \leq C 2^{2j\nu}.$$

*Proof.* Suppose the family $\psi_{j,\eta_k}$ is a frame (not necessarily tight). As the elements of a frame are bounded and the set $\{i,\ \psi^i_{jk} \neq 0\}$ is included in the set $\{C_1 2^j, \ldots, C_2 2^j\}$, we have

$$\sum_i \left[\frac{\psi^i_{jk}}{b_i}\right]^2 \leq C 2^{j\nu} \|\psi_{j,\eta_k}\|^2 \leq C' 2^{j\nu}. \qquad \square$$

**4.2. Convergence results in the Jacobi case.** The following theorem is the analogous of Theorem 2.1 in this case. As can be seen, the results there are at the same time more difficult to obtain (the following theorem does not cover the same range as the previous one) and richer since they furnish new rates of convergence.

**Theorem 4.2.** *Suppose that we are in the Jacobi case as stated above* ($\alpha \geq \beta > -\frac{1}{2}$). *We put*

$$t_\varepsilon = \varepsilon\sqrt{\log 1/\varepsilon},$$
$$2^J = t_\varepsilon^{-\frac{2}{1+2\nu}},$$

*choose* $\kappa \geq 16p\big[1 + \big(\frac{\alpha}{2} - \frac{\alpha+1}{p}\big)_+\big]$, *and suppose that we are in the regular case, i.e.*

$$b_i \sim i^{-\nu}, \quad \nu > -\frac{1}{2}.$$

*Then, if* $f = \sum_j \sum_k \beta_{j,\eta_k}\psi_{j,\eta_k}$ *is such that*

$$\Big(\sum |\beta_{j,\eta_k}|^p \|\psi_{j,\eta_k}\|_p^p\Big)^{1/p} \leq \rho_j 2^{-js}, \quad (\rho_j) \in l_r,$$

*it follows that*

$$\mathbb{E}\|\hat{f} - f\|_p^p \leq C[\log(1/\varepsilon)]^{p-1}\big[\varepsilon\sqrt{\log(1/\varepsilon)}\,\big]^{\mu p}$$

*with*

1. *if* $p < 2 + \frac{1}{\alpha+1/2}$, *then*
$$\mu = \frac{s}{s + \nu + \frac{1}{2}};$$

2. *if* $p > 2 + \frac{1}{\alpha+1/2}$, *then*
$$\mu = \frac{s}{s + \nu + \alpha + 1 - \frac{2(1+\alpha)}{p}}.$$

This theorem is proved in Kerkyacharian et al. [26]. Simulation results on these methods are given there, showing that their performances are far above the usual SVD methods in several cases. It is interesting to notice that the rates of convergence which are obtained here agree with the minimax rates evaluated in Johnstone and Silverman [23] where the case $p = 2$ is considered. But the second case ($p > 2 + \frac{1}{\alpha+1/2}$) shows a rate of convergence which is new in the literature. In [26], where the whole range of Besov bodies is considered, more atypical rates are given.

## 5. Direct models ($K = I$): a memento

**5.1. The density model.** The most famous nonparametric model consists in observing $n$ i.i.d. random variables having a common density $f$ on the interval $[0, 1]$, and in trying to give an estimation of $f$.

A standard route to perform this estimation consists in expanding the density $f$ in an orthonormal basis $\{e_k,\ k \in \mathbb{N}\}$ of a Hilbert space $\mathbb{H}$ – assuming implicitly that $f$ belongs to $\mathbb{H}$:

$$f = \sum_{l \in \mathbb{N}} \theta_l e_l.$$

If $\mathbb{H}$ happens to be the space $\mathbb{L}_2 = \{g :\ [0, 1] \mapsto \mathbb{R},\ \|g\|_2^2 := \int_0^1 g^2 < \infty\}$, we observe that

$$\theta_l = \int_0^1 e_l(x) f(x) dx = \mathbb{E} e_l(X_i).$$

Replacing the expectation by the empirical one leads to a standard estimate for $\theta_l$:

$$\hat{\theta}_l = \frac{1}{n} \sum_{i=1}^{n} e_l(X_i).$$

At this step, the simplest choice of estimate for $f$ is obviously:

$$\hat{f}_m = \sum_{i=1}^{m} \hat{\theta}_l e_l. \tag{18}$$

**5.2. From the density to the white noise model.** Before analysing the properties of the estimator defined above, let us observe that the previous approach (representing $f$ by its coefficients $\{\theta_k,\ k \geq 0\}$, leads to summarize the information in the following sequence model:

$$\{\hat{\theta}_k,\ k \geq 0\}. \tag{19}$$

We can write $\hat{\theta}_k =: \theta_k + u_k$, with

$$u_k = \frac{1}{n} \sum_{i=1}^{n} [e_k(X_i) - \theta_k],$$

The central limit theorem is a relatively convincing argument that the model (19) may be approximated by the following one:

$$\left\{\hat{\theta}_k = \theta_k + \frac{\eta_k}{\sqrt{n}},\ k \geq 0\right\}, \tag{20}$$

where the $\eta_k$'s are forming a sequence of i.i.d. gaussian, centered variables with fixed variance $\sigma^2$, say. Such an approximation requires more delicate calculations than these quick arguments and is rigourously proved in Nussbaum [35], see also Brown

and Low [5]. This model is the sequence space model associated to the following global observation, the so-called white noise model (with $\varepsilon = n^{-1/2}$):

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t \in [0, 1],$$

where for any $\varphi \in \mathbb{L}^2([0, 1], dt)$, $\int_{[0,1]} \varphi(t)dY_t = \int_{[0,1]} f(t)\varphi(t)dt + \varepsilon \int_{[0,1]} \varphi(t)dW_t$ is observable.

(20) formally consists in considering all the observables obtained for $\varphi = e_k$ for all $k$ in $\mathbb{N}$. Among nonparametric situations, the white noise model considered above is one of the simplest, at least technically. Mostly for this reason, this model has been given a central place in statistics, particularly by the Russian school, following Ibraguimov and Has'minskii (see for instance their book [20]). However it arises as an appropriate large sample limit to more general nonparametric models, such as regression with random design, or non independent spectrum estimation, diffusion models – see for instance [21], [4],. . . .

**5.3. The linear estimation: how to choose the tuning parameter *m*?** In (18), the choice of $m$ is crucial.

To better understand the situation let us have a look at the risk of the strategy $\hat{f}_m$. If we consider that, when deciding $\hat{f}_m$ when $f$ is the truth, we have a loss of order $\|\hat{f}_m - f\|_2^2$, then our risk will be the following mathematical expectation:

$$\mathbb{E}\|\hat{f}_m - f\|_2^2.$$

Of course this way of measuring our risk is arguable since there is no particular reason for the $\mathbb{L}_2$ norm to reflect well the features we want to recover in the signal. For instance, an $\mathbb{L}_\infty$-norm could be preferred because it is easier to visualize. In general, several $\mathbb{L}_p$ norms are considered (as it is the case in Sections 2.1 and 4.2). Here we restrict to the $\mathbb{L}_2$ case for sake of simplicity.

To avoid technical difficulties, we set ourselves in the case of a white noise model, considering that we observe the sequence defined in (20). Hence,

$$\mathbb{E}(\hat{\theta}_l - \theta_l)^2 = \frac{1}{n}\int_0^1 e_l(x)^2 dx = \frac{1}{n} := \varepsilon^2.$$

We are now able to obtain

$$\mathbb{E}\|\hat{f}_m - f\|_2^2 = \sum_{l \le m}(\hat{\theta}_l - \theta_l)^2 + \sum_{l>m}\theta_l^2 \le m\varepsilon^2 + \sum_{l>m}\theta_l^2.$$

Now assume that $f$ belongs to the following specified compact set of $l_2$:

$$\sum_{l>k}\theta_l^2 \le Mk^{-2s} \quad \text{for all } k \in \mathbb{N}_*, \tag{21}$$

for some $s > 0$ which is here an index of regularity directly connected to the size of the compact set in $l_2$ containing the function $f$. Then we obtain

$$\mathbb{E}\|\hat{f}_m - f\|_2^2 \leq m\varepsilon^2 + Mm^{-2s}.$$

We observe that the RHS is the sum of two factors: one (called the stochastic term) is increasing in $m$ and reflects the fact that because of the noise, the more coefficients we have to estimate, the larger the global error will be. The second one (called the bias term or approximation term) does not depend on the noise and is decreasing in $m$. The RHS is optimised by choosing $m = m_*(s) =: c(s, M)\varepsilon^{\frac{-2}{1+2s}}$. Then

$$\mathbb{E}\|\hat{f}_{m_*(s)} - f\|_2 \leq c'(s, M)\varepsilon^{\frac{-4s}{1+2s}}.$$

Let us observe that the more $f$ is supposed to be regular (in the sense the larger $s$ is), the less coefficients we need to estimate: a very irregular function ($s$ close to 0) requires almost as much as $\varepsilon^{-2} = n$ coefficients, which corresponds to estimate as many coefficients as the number of available observations – in the density model for instance. The rate obtained in (5.3) can be proved to be optimal in the following sense (minimax): if we consider the best estimator in its worst performance over the class of functions verifying (21), this estimator attains a rate of convergence which is (up to a constant) the one given in (5.3). See Tsybakov [40] for a detailed review of the minimax point of view.

**5.4. The thresholding estimation.** Let us now suppose that the constant $s$, which plays an essential role in the construction of the previous estimator is not known. This is realistic, since it is extremely rare to know in advance that the function we are seeking has a specified regularity. Also, the previous approach takes very seriously into account the order in which the basis is taken. Let us now present a very elegant way of addressing at the same time both of these issues. The thresholding techniques which have been known for long by engineers in electronic and telecommunications, was introduced in statistics in Donoho and Johnstone [14] and later in a series of papers on wavelet thresholding [12], [13]. It allies numerical simplicity to asymptotic optimality.

It starts from a different kind of observation. Let us introduce the following estimate:

$$\widetilde{f} = \sum_{k=0}^{B} \hat{\theta}_k \mathbb{I}\{|\hat{\theta}_k| \geq \kappa t_\varepsilon\} e_k. \tag{22}$$

Here the point of view is the following. We choose $B$ very large (i.e. almost corresponding to $s = 0$):

$$B = \varepsilon^{-2}\log 1/\varepsilon.$$

But instead of keeping all the coefficients $\theta_k$ such that $k$ is between 0 and $B$, we decide to kill those which are not above the threshold $t_\varepsilon$. The intuitive justification of

this choice is as follows. Assuming that $f$ has some kind of regularity condition like (21) (unknown, but real...), essentially means that the coefficients $\theta_k$ of $f$ are of small magnitude except perhaps a small number of them. Obviously, in the reconstruction of $f$, only the large coefficients will be significant. $t_\varepsilon$ is chosen in such a way that the noise $\hat{\theta}_k - \theta_k$ due to the randomness of the observation might be neglected:

$$t_\varepsilon = \varepsilon[\log 1/\varepsilon]^{-1/2}.$$

Now let us assume another type of condition on $f$ – easily interpreted by the fact that $f$ is sparsely represented in the basis $(e_k)$ – namely: there exists a positive constant $0 < q < 2$ such that

$$\sup_{\lambda>0} \lambda^q \#\{k,\ |\theta_k| \geq \lambda\} \leq M \quad \text{for all } k \in \mathbb{N}_*, \tag{23}$$

$$\begin{aligned}
\mathbb{E}\|\widetilde{f} - f\|_2^2 &= \sum_{l\leq B}(\hat{\theta}_l \mathbb{I}\{|\hat{\theta}_l| \geq \kappa t_\varepsilon\} - \theta_l)^2 + \sum_{l>B}\theta_l^2 \\
&\leq \sum_l (\hat{\theta}_l - \theta_l)^2 \mathbb{I}\{|\theta_l| \geq \kappa t_\varepsilon/2\} + \sum_l \theta_l^2 \mathbb{I}\{|\theta_l| \leq 2\kappa t_\varepsilon\} \\
&\quad + \sum_{l\leq B}[(\hat{\theta}_l - \theta_l)^2 + \theta_l^2]\mathbb{I}\{|\hat{\theta}_l - \theta_l| \geq \kappa t_\varepsilon/2\} + \sum_{l>B}\theta_l^2.
\end{aligned}$$

Now, using the probabilistic bounds

$$\mathbb{E}(\hat{\theta}_l - \theta_l)^2 = \varepsilon^2, \qquad \mathbb{P}(|\hat{\theta}_l - \theta_l| \geq \lambda) \leq 2\exp -\frac{\lambda^2}{2\varepsilon^2} \quad \text{for all } \lambda > 0,$$

and the fact that condition (23) implies

$$\sum_l \theta_l^2 \mathbb{I}\{|\theta_l| \leq 2\kappa t_\varepsilon\} \leq C t_\varepsilon^{2-q},$$

we get

$$\mathbb{E}\|\widetilde{f} - f\|_2^2 \leq M\varepsilon^2 t_\varepsilon^{-q} + C' t_\varepsilon^{2-q} + \varepsilon^{\kappa^2/8} B + \sum_{l>B}\theta_l^2.$$

It remains now to choose $\kappa^2 \geq 32$ in order to get

$$\mathbb{E}\|\widetilde{f} - f\|_2^2 \leq C' t_\varepsilon^{2-q} + \sum_{l>B}\theta_l^2,$$

and if we assume in addition to (23) that

$$\sum_{l>k}\theta_l^2 \leq M k^{-\frac{2-q}{2}} \quad \text{for all } k \in \mathbb{N}_*, \tag{24}$$

then we get

$$\mathbb{E}\|\widetilde{f} - f\|_2^2 \leq C'' t_\varepsilon^{2-q}$$

Note that the interesting point in this construction is that the regularity conditions imposed on the function $f$ are *not known* by the statistician, since they do not enter into the construction of the procedure. This property is called adaptation.

Now, to compare with the previous section, let us take $q = \frac{2}{1+2s}$. It is not difficult to prove that as soon as $f$ verifies (21), it automatically verifies (23) and (24). Hence $\widetilde{f}$ and $\hat{f}_{m*(s)}$ have the same rate of convergence up to a logarithmic term. If we neglect this logarithmic loss, we substantially gain here the fact that we need not know the apriori regularity conditions on the aim function. It can also be proved that in fact conditions (23) and (24) are defining a set which is substantially larger than the set defined by condition (21): for instance its entropy is strictly larger (see [28]).

## 6. Appendix: Proof of Theorem 2.1

In this proof, $C$ will denote an absolute constant which may change from one line to the other.

We can always suppose $p \geq \pi$. Indeed, if $\pi \geq p$ it is very simple to see that $B^s_{\pi,r}(M)$ is included into $B^s_{p,r}(M)$: as $2^{j[s+\frac{1}{2}-\frac{1}{p}]}\|\beta_{j\cdot}\|_{l_p} \leq 2^{j[s+\frac{1}{2}-\frac{1}{\pi}]}\|\beta_{j\cdot}\|_{l_\pi}$ (since $\chi_j$ is of cardinality $2^j$).

First we have the following decomposition:

$$\mathbb{E}\|\hat{f} - f\|_p^p \leq 2^{p-1}\left\{\mathbb{E}\Big\|\sum_{j=-1}^{J}\sum_{k\in\chi_j}(t(\hat{\beta}_{jk}) - \beta_{jk})\psi_{jk}\Big\|_p^p + \Big\|\sum_{j>J}\sum_{k\in\chi_j}\beta_{jk}\psi_{jk}\Big\|_p^p\right\}$$
$$=: \mathrm{I} + \mathrm{II}.$$

The term II is easy to analyse: since $f$ belongs to $B^s_{\pi,r}(M)$, using standard embedding results (which in this case simply follows from direct comparisons between $l_q$ norms) we have that $f$ also belong to $B^{s-(\frac{1}{\pi}-\frac{1}{p})_+}_{p,r}(M')$, for some constant $M'$. Hence

$$\Big\|\sum_{j>J}\sum_{k\in\chi_j}\beta_{jk}\psi_{jk}\Big\|_p \leq C2^{-J\left[s-(\frac{1}{\pi}-\frac{1}{p})_+\right]}.$$

Then we only need to verify that $\frac{s-(\frac{1}{\pi}-\frac{1}{p})_+}{1+2\nu}$ is always larger that $\alpha$, which is not difficult.

Bounding the term I is more involved. Using the triangular inequality together

with Hölder's inequality and property (8) for the second line, we get

$$\mathrm{I} \le 2^{p-1} J^{p-1} \sum_{j=-1}^{J} \mathbb{E} \Big\| \sum_{k\in\chi_j} (t(\hat\beta_{jk}) - \beta_{jk})\psi_{jk} \Big\|_p^p$$
$$\le 2^{p-1} J^{p-1} D_p \sum_{j=-1}^{J} \sum_{k\in\chi_j} \mathbb{E}|t(\hat\beta_{jk}) - \beta_{jk}|^p \|\psi_{jk}\|_p^p.$$

Now, we separate four cases:

$$\sum_{j=-1}^{J} \sum_{k\in\chi_j} \mathbb{E}|t(\hat\beta_{jk}) - \beta_{jk}|^p \|\psi_{jk}\|_p^p$$
$$= \sum_{j=-1}^{J} \sum_{k\in\chi_j} \mathbb{E}|t(\hat\beta_{jk}) - \beta_{jk}|^p \|\psi_{jk}\|_p^p \big\{ I\{|\hat\beta_{jk}| \ge \kappa t_\varepsilon \sigma_j\} + I\{|\hat\beta_{jk}| < \kappa t_\varepsilon \sigma_j\}\big\}$$
$$\le \sum_{j=-1}^{J} \sum_{k\in\chi_j} \Big[ \mathbb{E}|\hat\beta_{jk} - \beta_{jk}|^p \|\psi_{jk}\|_p^p I\{|\hat\beta_{jk}| \ge \kappa t_\varepsilon \sigma_j\}$$
$$\Big\{ I\Big\{|\beta_{jk}| \ge \frac{\kappa}{2} t_\varepsilon \sigma_j\Big\} + I\Big\{|\beta_{jk}| < \frac{\kappa}{2} t_\varepsilon \sigma_j\Big\}\Big\}$$
$$+ |\beta_{jk}|^p \|\psi_{jk}\|_p^p I\{|\hat\beta_{jk}| \le \kappa t_\varepsilon \sigma_j\}$$
$$\big\{ I\{|\beta_{jk}| \ge 2\kappa t_\varepsilon \sigma_j\} + I\{|\beta_{jk}| < 2\kappa t_\varepsilon \sigma_j\}\big\}\Big]$$
$$\le: Bb + Bs + Sb + Ss.$$

Notice that $\hat\beta_{jk} - \beta_{jk} = \sum_i \frac{Y_i - b_i f_i}{b_i} \psi_{jk}^i = \varepsilon \sum_i \xi_i \frac{\psi_{jk}^i}{b_i}$ is a centered gaussian random variable with variance $\varepsilon^2 \sum_i \big[\frac{\psi_{jk}^i}{b_i}\big]^2$. Also recall that we set $\sigma_j^2 =: \sum_i [\frac{\psi_{jk}^i}{b_i}]^2 \le C 2^{2j\nu}$ and denote by $s_q$ the $q$th absolute moment of the gaussian distribution when centered and with variance 1. Then, using standard properties of the gaussian distribution, for any $q \ge 1$ we have

$$\mathbb{E}|\hat\beta_{jk} - \beta_{jk}|^q \le s_q \sigma_j^q \varepsilon^q, \quad \mathbb{P}\{|\hat\beta_{jk} - \beta_{jk}| \ge \frac{\kappa}{2} t_\varepsilon \sigma_j\} \le 2\varepsilon^{\kappa^2/8}.$$

Hence

$$Bb \le \sum_{j=-1}^{J} \sum_{k\in\chi_j} s_p \sigma_j^p \varepsilon^p \|\psi_{jk}\|_p^p I\{|\beta_{jk}| \ge \frac{\kappa}{2} t_\varepsilon \sigma_j\},$$
$$Ss \le \sum_{j=-1}^{J} \sum_{k\in\chi_j} |\beta_{jk}|^p \|\psi_{jk}\|_p^p I\{|\beta_{jk}| < 2\kappa t_\varepsilon \sigma_j\}$$

and

$$
\begin{aligned}
Bs &\le \sum_{j=-1}^{J}\sum_{k\in\chi_j}[\mathbb{E}|\hat\beta_{jk}-\beta_{jk}|^{2p}]^{1/2}\Big[\mathbb{P}\Big\{|\hat\beta_{jk}-\beta_{jk}|\ge\frac{\kappa}{2}t_\varepsilon\sigma_j\Big\}\Big]^{1/2}\\
&\qquad\qquad \|\psi_{jk}\|_p^p I\Big\{|\beta_{jk}|<\frac{\kappa}{2}t_\varepsilon\sigma_j\Big\}\\
&\le \sum_{j=-1}^{J}\sum_{k\in\chi_j} s_{2p}^{1/2}\sigma_j^p\varepsilon^p 2^{1/2}\varepsilon^{\kappa^2/16}\|\psi_{jk}\|_p^p I\Big\{|\beta_{jk}|<\frac{\kappa}{2}t_\varepsilon\sigma_j\Big\}\\
&\le C\sum_{j=-1}^{J}2^{jp(\nu+\frac12)}\varepsilon^p\varepsilon^{\kappa^2/16}\le C\varepsilon^{\kappa^2/16}.
\end{aligned}
$$

Now, if we remark that the $\beta_{jk}$'s are necessarily all bounded by some constant (depending on $M$) since $f$ belongs to $B^s_{\pi,r}(M)$, and using (7),

$$
\begin{aligned}
Sb &\le \sum_{j=-1}^{J}\sum_{k\in\chi_j}|\beta_{jk}|^p\|\psi_{jk}\|_p^p\mathbb{P}\{|\hat\beta_{jk}-\beta_{jk}|\ge 2\kappa t_\varepsilon\sigma_j\}I\{|\beta_{jk}|\ge 2\kappa t_\varepsilon\sigma_j\}\\
&\le \sum_{j=-1}^{J}\sum_{k\in\chi_j}|\beta_{jk}|^p\|\psi_{jk}\|_p^p 2\varepsilon^{\kappa^2/8}I\{|\beta_{jk}|\ge 2\kappa t_\varepsilon\sigma_j\}\\
&\le C\sum_{j=-1}^{J}2^{j\frac{p}{2}}\varepsilon^{\kappa^2/8}\le C\varepsilon^{\frac{\kappa^2}{8}-\frac{p}{2(2\nu+1)}}.
\end{aligned}
$$

It is easy to check that in all cases, if $\kappa^2\ge 16p$ the terms $Bs$ and $Sb$ are smaller than the rates given in the theorem.

Using (7) and condition (10), for any $z\ge 0$ we have

$$
\begin{aligned}
Bb &\le C\varepsilon^p\sum_{j=-1}^{J}2^{j(\nu p+\frac{p}{2}-1)}\sum_{k\in\chi_j}I\Big\{|\beta_{jk}|\ge\frac{\kappa}{2}t_\varepsilon\sigma_j\Big\}\\
&\le C\varepsilon^p\sum_{j=-1}^{J}2^{j(\nu p+\frac{p}{2}-1)}\sum_{k\in\chi_j}|\beta_{jk}|^z[t_\varepsilon\sigma_j]^{-z}\\
&\le Ct_\varepsilon{}^{p-z}\sum_{j=-1}^{J}2^{j[\nu(p-z)+\frac{p}{2}-1]}\sum_{k\in\chi_j}|\beta_{jk}|^z.
\end{aligned}
$$

Also, for any $p \geq z \geq 0$,

$$Ss \leq C \sum_{j=-1}^{J} 2^{j(\frac{p}{2}-1)} \sum_{k\in\chi_j} |\beta_{jk}|^z \sigma_j^{p-z} [t_\varepsilon]^{p-z}$$
$$\leq C[t_\varepsilon]^{p-z} \sum_{j=-1}^{J} 2^{j(\nu(p-z)+\frac{p}{2}-1)} \sum_{k\in\chi_j} |\beta_{jk}|^z.$$

So in both cases we have the same bound to investigate. We will write this bound in the following form (forgetting the constant):

$$\mathrm{I} + \mathrm{II} = {t_\varepsilon}^{p-z_1} \Big[ \sum_{j=-1}^{j_0} 2^{j[\nu(p-z_1)+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{z_1} \Big]$$
$$+ {t_\varepsilon}^{p-z_2} \Big[ \sum_{j=j_0+1}^{J} 2^{j[\nu(p-z_2)+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{z_2} \Big].$$

The constants $z_i$ and $j_0$ will be chosen depending on the cases.

Let us first consider the case where $s \geq \big(\nu + \frac{1}{2}\big)\big(\frac{p}{\pi} - 1\big)$. Put

$$q = \frac{p(2\nu+1)}{2(s+\nu)+1}$$

and observe that, on the considered domain, $q \leq \pi$ and $p > q$. In the sequel it will be used that we automatically have $s = \big(\nu + \frac{1}{2}\big)\big(\frac{p}{q} - 1\big)$. Taking $z_2 = \pi$ we get

$$\mathrm{II} \leq {t_\varepsilon}^{p-\pi} \Big[ \sum_{j=j_0+1}^{J} 2^{j[\nu(p-\pi)+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{\pi} \Big].$$

Now, as

$$\frac{p}{2q} - \frac{1}{\pi} + \nu\left(\frac{p}{q} - 1\right) = s + \frac{1}{2} - \frac{1}{\pi}$$

and

$$\sum_{k\in\chi_j} |\beta_{jk}|^{\pi} = 2^{-j(s+\frac{1}{2}-\frac{1}{\pi})} \tau_j$$

with $(\tau_j)_j \in l_r$ (this is a consequence of the fact that $f \in B^s_{\pi,r}(M)$ and (6)), we can write

$$\mathrm{II} \leq {t_\varepsilon}^{p-\pi} \sum_{j=j_0+1} 2^{jp(1-\frac{\pi}{q})(\nu+\frac{1}{2})} \tau_j^{\pi}$$
$$\leq C {t_\varepsilon}^{p-\pi} 2^{j_0 p(1-\frac{\pi}{q})(\nu+\frac{1}{2})}.$$

The last inequality is true for any $r \geq 1$ if $\pi > q$ and for $r \leq \pi$ if $\pi = q$. Notice that $\pi = q$ is equivalent to $s = (\nu + \frac{1}{2})(\frac{p}{\pi} - 1)$. Now if we choose $j_0$ such that $2^{j_0 \frac{p}{q}(\nu+\frac{1}{2})} \sim {t_\varepsilon}^{-1}$ we get the bound

$$ {t_\varepsilon}^{p-q} $$

which exactly gives the rate asserted in the theorem for this case.

As for the first part of the sum (before $j_0$), we have, taking now $z_1 = \widetilde{q}$, with $\widetilde{q} \leq \pi$, so that $\big[\frac{1}{2^j}\sum_{k\in\chi_j} |\beta_{jk}|^{\widetilde{q}}\big]^{\frac{1}{\widetilde{q}}} \leq \big[\frac{1}{2^j}\sum_{k\in\chi_j} |\beta_{jk}|^{\pi}\big]^{\frac{1}{\pi}}$, and using again (6),

$$\begin{aligned}
\mathrm{I} &\leq {t_\varepsilon}^{p-\widetilde{q}}\Big[\sum_{-1}^{j_0} 2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{\widetilde{q}}\Big] \\
&\leq {t_\varepsilon}^{p-\widetilde{q}}\Big[\sum_{-1}^{j_0} 2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-\frac{\widetilde{q}}{\pi}]} \sum_{k\in\chi_j} |\beta_{jk}|^{\pi}]^{\frac{\widetilde{q}}{\pi}}\Big] \\
&\leq {t_\varepsilon}^{p-\widetilde{q}} \sum_{-1}^{j_0} 2^{j[(\nu+\frac{1}{2})p(1-\frac{\widetilde{q}}{q})]} \tau_j^{\widetilde{q}} \\
&\leq C {t_\varepsilon}^{p-\widetilde{q}} 2^{j_0[(\nu+\frac{1}{2})p(1-\frac{\widetilde{q}}{q})]} \\
&\leq C {t_\varepsilon}^{p-q}.
\end{aligned}$$

The last two lines are valid if $\widetilde{q}$ is chosen strictly smaller than $q$ (this is possible since $\pi \geq q$).

Let us now consider the case where $s < \big(\nu + \frac{1}{2}\big)\big(\frac{p}{q} - 1\big)$, and choose

$$ q = \frac{p}{2\big(s + \nu - \frac{1}{\pi}\big) + 1} $$

in such a way that we easily verify that $p - q = 2\frac{s-1/\pi+1/p}{1+2(\nu+s-1/\pi)}$, $q - \pi = \frac{(p-\pi)(1+2\nu)}{2\big(s+\nu-\frac{1}{\pi}\big)+1} > 0$, because $s$ is supposed to be larger than $\frac{1}{\pi}$. Furthermore we also have $s + \frac{1}{2} - \frac{1}{\pi} = \frac{p}{2q} - \frac{1}{q} + \nu\big(\frac{p}{q} - 1\big)$.

Hence taking $z_1 = \pi$ and using again the fact that $f$ belongs to $B^s_{\pi,r}(M)$,

$$\begin{aligned}
\mathrm{I} &\leq {t_\varepsilon}^{p-\pi}\Big[\sum_{-1}^{j_0} 2^{j[\nu(p-\pi)+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{\pi}\Big] \\
&\leq {t_\varepsilon}^{p-\pi} \sum_{-1}^{j_0} 2^{j[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\pi)]} \tau_j^{\pi} \\
&\leq C {t_\varepsilon}^{p-\pi} 2^{j_0[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\pi)]}.
\end{aligned}$$

This is true since $\nu + \frac{1}{2} - \frac{1}{p}$ is also strictly positive because of our constraints. If we now take $2^{j_0 \frac{p}{q}(\nu+\frac{1}{2}-\frac{1}{p})} \sim {t_\varepsilon}^{-1}$ we get the bound

$$ {t_\varepsilon}^{p-q} $$

which is the rate stated in the theorem for this case.

Again, for II, we have, taking now $z_2 = \widetilde{q} > q \ (> \pi)$,

$$\begin{aligned}
\mathrm{II} &\le {t_\varepsilon}^{p-\widetilde{q}} \Big[ \sum_{j=j_0+1}^{J} 2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-1]} \sum_{k\in\chi_j} |\beta_{jk}|^{\widetilde{q}} \Big] \\
&\le C {t_\varepsilon}^{p-\widetilde{q}} \sum_{j=j_0+1} 2^{j[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\widetilde{q})]} z_j^{\frac{\widetilde{q}}{\pi}} \\
&\le C {t_\varepsilon}^{p-\widetilde{q}} 2^{j_0[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\widetilde{q})]} \\
&\le C {t_\varepsilon}^{p-q}.
\end{aligned}$$

## References


[1] Abramovich, F., and Silverman, B. W., Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85** (1) (1998), 115–129 .

[2] Antoniadis, A., and Bigot, J., Poisson inverse models. Preprint, Grenoble 2004.

[3] Antoniadis, A., Fan,J., and Gijbels, I., A wavelet method for unfolding sphere size distributions. *Canad. J. Statist.* **29** (2001), 251–268.

[4] Brown, Lawrence D., Cai, T. Tony, Low, Mark G., and Zhang, Cun-Hui Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** (3) (2002), 688–707.

[5] Brown, Lawrence D., and Low, Mark G., Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** (6) (1996), 2384–2398.

[6] Cavalier, Laurent, and Tsybakov, Alexandre, Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123** (3) (2002), 323–354.

[7] Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov., A. B., Oracle inequalities for inverse problems. *Ann. Statist.* **30** (3) (2002), 843–874.

[8] Cohen, Albert, Hoffmann, Marc, and Reiß, Markus, Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.* **42** (4) (2004), 1479–1501 (electronic).

[9] Cruz-Orive, L. M., Distribution-free estimation of sphere size distributions from slabs showing overprojections and truncations, with a review of previous methods. *J. Microscopy* **131** (1983), 265–290.

[10] Dicken, V., and Maass, P., Wavelet-Galerkin methods for ill-posed problems. *J. Inverse Ill-Posed Probl.* **4** (3) (1996), 203–221.

[11] Donoho, David L., Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** (2) (1995), 101–126.

[12] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D., Wavelet shrinkage: Asymptopia? *J. Royal Statist. Soc. Ser. B* **57** (1995), 301–369.

[13] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D., Density estimation by wavelet thresholding. *Ann. Statist.* **24** (1996), 508–539.

[14] Donoho, D. L., Johnstone, I. M., Minimax risk over $\ell_p$-balls for $\ell_q$-error. *Probab. Theory Related Fields* **99** (1994), 277–303.

[15] Duffin, R. J., and Schaeffer, A. C., A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.* **72** (1952), 341–366.

[16] Efromovich, Sam, and Koltchinskii, Vladimir, On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** (7) (2001), 2876–2894.

[17] Fan, J., and Koo, J. K., Wavelet deconvolution. *IEEE Trans. Inform. Theory* **48** (3) (2002), 734–747.

[18] Frazier, M., Jawerth, B., and Weiss, G., *Littlewood-Paley theory and the study of function spaces.* CBMS Reg. Conf. Ser. Math. 79, Amer. Math. Soc., Providence, RI, 1991.

[19] Goldenshluger, Alexander, and Pereverzev, Sergei V., On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9** (5) (2003), 783–807.

[20] Ibragimov, I. A., and Hasminskii, R. Z., *Statistical estimation*. Appl. Math. 16, Springer-Verlag, New York 1981.

[21] Jähnisch, Michael, and Nussbaum, Michael, Asymptotic equivalence for a model of independent non identically distributed observations. *Statist. Decisions* **21** (3) (2003), 197–218.

[22] Johnstone, Iain M., Kerkyacharian, Gérard, Picard, Dominique, and Raimondo, Marc, Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** (3) (2004), 547–573.

[23] Johnstone, Iain M., and Silverman, Bernard W., Discretization effects in statistical inverse problems. *J. Complexity* **7** (1) (1991), 1–34.

[24] Kalifa, Jérôme, and Mallat, Stéphane, Thresholding estimators for linear inverse problems and deconvolutions. *Ann. Statist.* **31** (1) (2003), 58–109.

[25] Kerkyacharian, G., Petrushev, P., Picard, D., and Xu, Y., Localized polynomials and frames induced by laguerre functions. Preprint, 2005.

[26] Kerkyacharian, G., Picard, D., Petrushev, P., and Willer, T., Needvd: second generation wavelets for estimation in inverse problems. Preprint, LPMA 2006.

[27] Kerkyacharian, G., Picard, D., and Raimondo, M., Adaptive boxcar deconvolution on full lebesgue measure sets. Preprint, LPMA 2005.

[28] Kerkyacharian, G., and Picard, D., Thresholding algorithms and well-concentrated bases. *Test* **9** (2) (2000).

[29] Mair, Bernard A., and Ruymgaart, Frits H., Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56** (5) (1996), 1424–1444.

[30] Mallat, Stéphane, *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA, 1998.

[31] Mathé, Peter, and Pereverzev, Sergei V., Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems* **19** (3) (2003), 789–803.

[32] Meyer, Yves, *Ondelettes et opérateurs. I*. Actualités Mathématiques. Hermann, Paris 1990.

[33] Narcowich, F. J., Petrushev, P., and Ward, J. M., Localized tight frames on spheres. *SIAM J. Math. Anal.* **38** (2) (2006), 574–594.

[34] Neelamani, R., Choi, H., and Baranuik, R., Wavelet-based deconvolution for ill-conditioned systems. Preprint, 2000; http://www-dsp.rice.edu/publications/pub/neelsh98icassp.pdf.

[35] Nussbaum, Michael, Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** (6) (1996), 2399–2430.

[36] Nyshka, D., Wahba, G., Goldfarb, S., and Pugh, T., Cross validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross sections. *J. Amer. Statist. Assoc.* **79** (1984), 832–8464.

[37] Pensky, M., and Vidakovic, B., Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27** (1999), 2033–2053.

[38] Petrushev, P., and Xu, Y., Localized polynomials frames on the interval with jacobi weights. *J. Fourier Anal. Appl.* **11** (5) (2005), 557–575.

[39] Petrushev, P., and Xu, Y., Localized polynomials kernels and frames (needlets) on the ball. 2005. IMI 2005.

[40] Tsybakov, Alexandre B., *Introduction à l'estimation non-paramétrique.* Math. Appl. (Berlin) 41, Springer-Verlag, Berlin 2004.

[41] Tsybakov, Alexandre, On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (9) (2000), 835–840.

[42] Wicksell, S. D., The corpuscle problem: a mathematical study of a biometric problem. *Biometrika* **17** (1925), 84–99.

[43] Willer, T., Deconvolution in white noise with a random blurring effect. Preprint, LPMA 2005.

LPMA, Universités Paris VII et Paris X, CNRS, 175 rue du Chevaleret, 75013 Paris, France

# Conformal restriction properties

Wendelin Werner*

**Abstract.** We give an introduction to some aspects of recent results concerning conformally invariant measures. We focus in this note on the conformal restriction properties of some measures on curves and loops in the plane, and see that these properties in fact almost characterize the measures and allow to classify them. For example, there basically exists a unique measure $\mu$ on the set of self-avoiding loops in the plane, such that for any two conformally equivalent domains $D$ and $D'$, the restrictions of $\mu$ to the set of loops remaining in $D$ and in $D'$ are conformally equivalent.

This enables to show that a priori different discrete models define the same curves in the scaling limit and exhibit some surprising symmetries. It gives also a way to tie links between these concrete measures on curves and conformal field theory. Important roles in this theory are played by Brownian loops and by the Schramm–Loewner Evolutions (SLE).

Most of the results described in this paper were derived in joint work with Greg Lawler, and Oded Schramm.

**Mathematics Subject Classification (2000).** Primary 60K35; Secondary 82B27, 60J65, 30Cxx.

**Keywords.** Conformal invariance, random curves, random loops, Brownian motion, percolation.

## 1. A very brief introduction

The last years have seen progress in the mathematical understanding of random two-dimensional structures arising as scaling limits of two-dimensional systems from statistical physics. These probabilistic questions are related to complex analysis considerations (because conformal invariance plays an important role in the description of these objects) and to conformal field theory (that had been developed by theoretical physicists precisely to understand these questions).

Mathematically speaking, one can broadly distinguish two types of questions: Firstly, proving the convergence of the natural discrete lattice-based models from statistical physics to conformally invariant scaling limits. This aspect based on specific lattice models will be discussed in Schramm's and Smirnov's papers in the present proceedings, and will not be the main focus of the present paper. The second type of questions is to define directly the possible continuous limiting objects and to study their

*The author acknowledges the support of the Institut Universitaire de France.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

properties. Two ideas have emerged and can be fruitfully combined to study these continuous objects: The Schramm–Loewner Evolutions (SLE) are random planar curves that are explicitly defined via iterations of random conformal maps, and they appear to be the only ones that combine conformal invariance with a certain Markov property. This shows that they are the only possible conformally invariant scaling limits of interfaces of the critical lattice models. Another instrumental idea is to study how close or how different the random objects defined in different but close domains are, and to see what the conformally invariant possibilities are. This very last approach will be the main focus of the present survey. We warn the reader that we will here remain on a rather general introductory level.

## 2. Conformal invariance of planar Brownian paths

In this section, we first recall Paul Lévy's result on conformal invariance of planar Brownian paths. We then describe some conformally invariant measures on Brownian loops and Brownian excursions.

**2.1. Paul Lévy's theorem.** Consider a simple random walk $(S_n, n \geq 0)$ on the square lattice $\mathbb{Z}^2$ (but in fact any planar lattice with some rotational symmetry would do) started from the origin (i.e. $S_0 = 0$). At each integer time, this random walk moves independently to one of its four neighbors with probability $1/4$. In other words, the probability that the first $n$ steps of $S$ are exactly a given nearest-neighbor path on the lattice is equal to $4^{-n}$. It is a simple consequence of the central limit theorem that when $N \to \infty$, the law of $(S_{[2Nt]}/\sqrt{N}, t \geq 0)$ converges in some suitable topology to that of a continuous random two-dimensional path $(B_t, t \geq 0)$ with Gaussian independent increments called planar Brownian motion.

It should be noted that planar Brownian paths have a rather complicated geometry. Even if their Lebesgue measure in the plane is almost surely equal to zero, the Hausdorff dimension of a Brownian path is equal to 2 (this can be related to the $\sqrt{N}$ normalization of the simple random walk). Also, there almost surely exists exceptional points of any (including infinite) multiplicity on planar Brownian paths (see [24] and the references therein).

Elementary properties of Gaussian random variables show that the law of the process $B$ is invariant under rotations in the plane, and that it is also scale-invariant (this is also quite clear from the normalization of the random walk) in the following sense: For each given $\lambda > 0$, the laws of $(B_{\lambda^2 t}, t \geq 0)$ and of $(\lambda B_t, t \geq 0)$ are identical. In other words, if one looks at the path of a Brownian motion with a magnifying glass, one sees exactly a Brownian motion, but running at a faster "speed". Paul Lévy (see e.g. [25]) has observed more than fifty years ago that planar Brownian paths exhibit conformal invariance properties that generalize scale-invariance and rotation-invariance, and that we are now describing:

Consider two given conformally equivalent planar domains $D$ and $D'$: These are two open subsets of $\mathbb{C}$ such that there exists an angle-preserving (and orientation-preserving) bijection (i.e. a conformal map) $\Phi$ from $D$ onto $D'$. Recall that when $D$ and $D'$ are two simply connected proper open subsets of the plane, then by Riemann's mapping Theorem, there exists a three-dimensional family of such conformal maps from $D$ onto $D'$. Consider a point $z$ in $D$ and define its image $z' = \Phi(z)$. Then, define a planar Brownian motion $(B_t, t \geq 0)$ that is started from $B_0 = z$, and denote by $T$ its exit time from the domain $D$ (i.e. $T = \inf\{t \geq 0 : B_t \notin D\}$). For each $t < T$, one can therefore define $\Phi(B_t)$, and when $t \to T$, $\Phi(B_t)$ hits the boundary of $D' = \Phi(D)$. Then:

**Theorem 2.1** (Paul Lévy). *The path $(\Phi(B_t), t \leq T)$ is a time-changed planar Brownian motion in $D'$, started at $z'$ and stopped at its first exit time of $D'$.*

The time-change means that there exists a (random continuous increasing) time-reparametrization $t = t(s)$ such that $(\Phi(B_{t(s)}), s \geq 0)$ is exactly a Brownian motion in $D'$. In order to state exact conformal invariance properties, we will from now on consider paths defined "modulo increasing time reparametrization".

Lévy's Theorem is nowadays usually viewed as a standard application of stochastic calculus (Itô's formula). It has led to probabilistic approaches to aspects of potential theory and complex analysis.

**2.2. Brownian excursions, Brownian loops.** It might be desirable to define conformally invariant random objects in a domain $D \subset \mathbb{C}$, but where no marked point in $D$ is given. In Lévy's Theorem, the starting point of the Brownian path is such a special prescribed point. There are (at least) two natural ways to get rid of it without loosing conformal invariance, that both give rise to infinite measures (i.e. measures with an infinite total mass) on Brownian curves.

- A first possibility, described in [22], is to consider Brownian paths that start and end at the boundary of $D$: Call an excursion in $D$ a continuous path $(e(t), 0 \leq t \leq T)$ such that $e(0, T) \subset D$ and $e(0) \in \partial D$, $e(T) \in \partial D$. Then, for each $D$, one can define an infinite measure $\mathrm{exc}_D$ on the set of "Brownian" excursions in $D$ (with unprescribed time-length) in such a way that the image under a conformal map $\Phi$ from $D$ onto $D'$ of the measure $\mathrm{exc}_D$ is identical to $\mathrm{exc}_{D'}$ modulo time-change.

  One way to describe the measure in the case where $D$ is equal to the unit disc $\mathbb{U}$ (and therefore in the case of all other simply connected domains $D$ via conformal invariance) is to take the limit when $\varepsilon$ goes to zero of $\varepsilon^{-1}$ times the law of a Brownian motion started uniformly on the circle $(1 - \varepsilon)\partial\mathbb{U}$ and stopped at its first hitting time of the unit circle $\partial\mathbb{U}$. One can also view these measures $\mathrm{exc}_D$ as the scaling limits (when $\delta \to 0$) of the measures on discrete excursions on approximations of $D$ by a subset of $\delta\mathbb{Z}^2$ that assign a mass $4^{-n}$ to each discrete excursion with $n$ steps (see e.g. [12] for precise estimates).

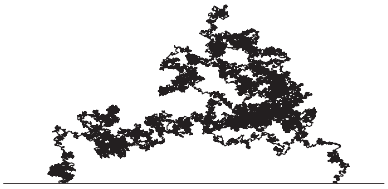


Figure 1. A Brownian excursion in the upper half-plane.

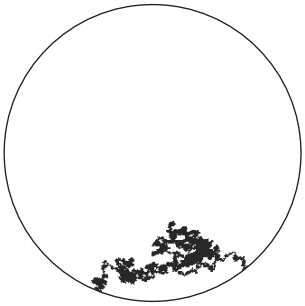


Figure 2. Its conformal image in the unit disc.

- A second possibility, described in [23], that will be important in the present paper, is to consider loops instead of open-ended paths. We say that a continuous planar path $(\ell_t, 0 \leq t \leq T)$ is a rooted loop if $\ell_0 = \ell_T$. The term rooted is used to emphasize that with this definition, there is a marked point on the loop, namely the starting point $\ell_0$. Note that it is possible to re-root a given loop by defining $(\ell'_t = \ell_{t+t_0},\ 0 \leq t \leq T)$ for a given fixed $t_0$ (where $\ell$ is extended into a $T$-periodic function). We may want to say however that $\ell$ and $\ell'$ define in fact the same unrooted loop. Hence, we call an unrooted loop the equivalence class of a rooted loop modulo the equivalence defined by this re-rooting procedure. In order to simplify the conformal invariance statements, we will also say that an unrooted loop is defined modulo increasing continuous time-reparametrizations.

  Then [23], there exists a measure $M$ on the set of unrooted (Brownian) loops in the plane with strong conformal invariance properties: For any two conformally equivalent open domains $D$ and $D' = \Phi(D)$, if $M_D$ (resp. $M'_D$) denotes the measure $M$ restricted to the set of loops that stay in $D$ (resp. $D'$), then the image measure of $M_D$ under the conformal map $\Phi$ from $D$ onto $D'$ is exactly the measure $M_{D'}$.

  One can view this measure $M$ as the limit when $\delta$ goes to zero of the measures on discrete unrooted loops in $\delta\mathbb{Z}^2$ that assign a mass $4^{-n}$ to each loop with $n$

Figure 3. A Brownian loop.

steps (see e.g. [20] for precise estimates). A direct construction of $M$ goes as follows ([23]): It is easy to define the law $P_{z,T}$ of a Brownian loop with a given time-length $T$ that starts and ends at a given point $z$. This can be viewed as the conditioning of a Brownian path $(B_t, t \leq T)$ started from $B_0 = z$ by the event $B_T = z$ (this event has zero probability but it is no big deal to make sense of this). Then, one can define a measure $\tilde{M}$ on rooted Brownian loops by integrating the starting point $z$ with respect to the Lebesgue measure in the plane, and the time-length by the measure $dT/T^2$. Then, $M$ is just the measure on the set of unrooted loops induced by $\tilde{M}$.

Note that $\mathrm{exc}_D$ and $M$ are infinite measures (this follows readily from the scale invariance of $M$ and from the scale-invariance of the excursion measure $\mathrm{exc}_{\mathbb{H}}$ in the upper half-plane), so that we can in both cases choose a normalization constant as we wish (i.e. multiply the measures by a well-chosen constant). In fact, the different descriptions of the measures that we did (and will) give differ by a multiplicative constant, and we will not really care here about the exact choice of the constant in the definition of $M$.

Since these are measures on Brownian paths, they are supported on the set of paths with Hausdorff dimension equal to two, but that the mass of the set of paths that go through any given prescribed point $z$ is equal to zero.

In a way, both these measures are invariant under a larger class of conformal transformations than the killed Brownian motions defined in the previous subsection because no marked starting point is prescribed. Just as killed Brownian motions describe conformally invariant quantities associated to a given marked point such as the harmonic measure, these two measures define also natural conformally invariant quantities that can be related to extremal distances or Schwarzian derivatives for

instance.

Let us finally define a further useful Brownian measure, the Brownian excursion measure with prescribed endpoints: The excursion measure $\mathrm{exc}_{\mathbb{U}}$ can be decomposed according to the starting and endpoints of the Brownian excursions. This gives rise for each $A \neq B$ on $\partial\mathbb{U}$ to a probability measure $e_{\mathbb{U},A,B}$ on the set of Brownian excursions from $A$ to $B$ in $\mathbb{U}$. This defines (not surprisingly) again a conformally invariant family of probability measures $(e_{D,A,B})$ where $(D, A, B)$ spans the set $\mathcal{T}$ the set of triplets $(D, A, B)$ such that $D$ is a simply connected proper subset of $\mathbb{C}$ and $A$ and $B$ denote two distinct prime ends of $D$. When the boundary of $D$ is a smooth self-avoiding loop, this means that $A$ and $B$ are two distinct boundary points. When $\Phi$ is a conformal map from $D$ onto $D'$, then "$\Phi(A)$" and "$\Phi(B)$" are then by definition distinct prime ends of $D' = \Phi(D)$.

## 3. Conformal restriction

We have so far defined some measures on Brownian paths with conformal invariance properties. This means that for each (simply connected) domain, we had a measure $m_D$ on paths in $D$, and that the family $(m_D)$ is conformally invariant (i.e. $\Phi \circ m_D = m_{\Phi(D)}$). But when $D' \subset D$, it is also natural to compare $m_{D'}$ with the measure $m_D$ restricted to those paths that stay in $D'$. The conformal restriction property basically requires that these two measures coincide (and that conformal invariance also holds).

**3.1. Loops.** Suppose that $\nu$ is a measure on loops in the plane. As in the rest of the paper, the loops are unrooted and defined modulo increasing time-reparametrizations. For each open domain $D$, we define $\nu_D$ to be the measure $\nu$ restricted to the set of loops that stay in $D$.

**Definition 3.1.** We say that $\nu$ satisfies conformal restriction (resp. conformal restriction for simply connected domains) if for any open domain (resp. open simply connected domain) $D$ and any conformal map $\Phi\colon D \to \Phi(D)$, one has $\Phi \circ \nu_D = \nu_{\Phi(D)}$.

We have already seen one measure satisfying conformal restriction in the previous section: The measure $M$ on Brownian loops in the plane.

Let us now describe a simple argument that shows that all measures that satisfy conformal restriction are closely related. Before that, let us introduce the notion of the filling of a loop. If $\gamma$ is a loop in the plane, we define its filling $K(\gamma)$ to be the complement of the unbounded connected component of $\mathbb{C} \setminus \gamma$. In other words, $K(\gamma)$ is obtained by filling in all the bounded connected components of the complement of $\gamma$. Clearly, any measure on loops defines a measure on their fillings, and we can also define the conformal restriction property for measures on fillings.

**Proposition 3.2** ([42]). *Up to multiplication by a positive constant, there exists a unique measure on fillings that satisfies conformal restriction for simply connected domains. It can be defined as the measure on filling of Brownian loops.*

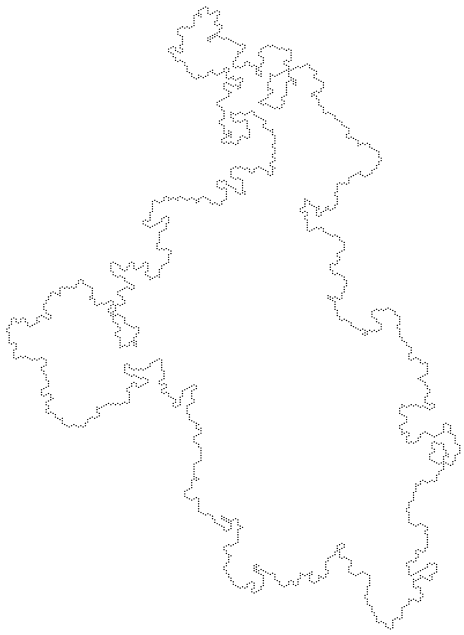


Figure 4. A self-avoiding loop.

*Proof* (*sketch*). In this proof, we will always discuss conformal restriction *for simply connected domains*. The existence part of the proposition follows from the fact that the measure $M$ on Brownian loops exists and satisfies conformal restriction (so that the fillings of Brownian loops satisfy conformal restriction as well). It remains to prove the uniqueness statement.

Consider the family $\mathcal{U}$ of conformal maps $\varphi$ from some (unprescribed) simply connected subset $U$ of the unit disc $\mathbb{U}$ containing the origin onto the unit disc $\mathbb{U}$, such that $\varphi(0) = 0$ and $\varphi'(0)$ is a positive real number. Riemann's mapping theorem shows that for any simply connected domain $U \subset \mathbb{U}$ with $0 \in U$, there exists a unique $\varphi = \varphi_U \in \mathcal{U}$ from $U$ onto $\mathbb{U}$. Note that $\mathcal{U}$ is closed under composition: If $\varphi_U$ and $\varphi_V$ are in $\mathcal{U}$, then so is $\psi = \varphi_U \circ \varphi_V$ (it is a conformal map from $\varphi_V^{-1} \circ \varphi_U^{-1}(\mathbb{U}) = \varphi_V^{-1}(U)$ onto $\mathbb{U}$ with the right properties at the origin). Note that of course, $\log \psi'(0) = \log \varphi_U'(0) + \log \varphi_V'(0)$. It is also straightforward to check that $\varphi_V'(0) \geq 1$ because $V \subset \mathbb{U}$.

Suppose now that a measure $\nu$ on fillings satisfies conformal restriction. Let us define for each $\varphi_U \in \mathcal{U}$,

$$A(\varphi_U) = \nu(\{K : 0 \in K,\ K \subset \mathbb{U},\ K \not\subset U\}).$$

This is the mass of fillings containing the origin, that stay in $\mathbb{U}$ but not in $U$. Then, it is easy to see that

$$A(\varphi_U \circ \varphi_V) = A(\varphi_U) + A(\varphi_V).$$

Indeed, there are two types of fillings that contain the origin, stay in $\mathbb{U}$ but not in $\varphi_V^{-1} \circ \varphi_U^{-1}(\mathbb{U})$:

- Those that do not stay in $V = \varphi_V^{-1}(\mathbb{U})$ and the set of these fillings has a $\nu$-mass equal to $A(\varphi_V)$ by definition.
- Those that stay in $V = \varphi_V^{-1}(\mathbb{U})$ but not in $\varphi_V^{-1}(\varphi_U^{-1})(\mathbb{U}) = \varphi_V^{-1}(U)$. But by conformal invariance (via the mapping $\varphi_V$), this set is conformally equivalent to the set of loops that stay in $\mathbb{U}$ and not in $U$. So, its $\nu$-mass is $A(\varphi_U)$.

Rather soft considerations (for instance involving Loewner's approximation of any mapping in $\mathcal{U}$ by iterations of slit mappings) then imply that the functional $A$ is necessarily of the form $A(\varphi_U) = c \log \varphi_U'(0)$ for a positive constant $c$.

Hence, it follows by conformal invariance that for each $z \in D' \subset D$, the $\nu$-mass of the set of fillings that contain $z$, stay in the simply connected domain $D$ but not in the simply connected domain $D'$ is equal to $c$ times the logarithm of the derivative at $z$ of the conformal map from $D'$ onto $D$ that fixes $z$ and has positive derivative at $z$. Soft arguments (of the type "a finite measure is characterized by its values on a intersection-stable set that generates the $\sigma$-field") then show that (for each choice of $c$) this characterizes the measure $\nu$ uniquely. This implies the uniqueness part of the proposition. □

It is possible to show that the boundary of a Brownian loop is almost surely a self-avoiding loop (the fact that it is a continuous loop is straightforward, but the fact that it has no double point requires some estimates, see e.g. [4]). Hence, the proposition shows that modulo multiplication by a positive constant, there is a unique measure $\mu$ on self-avoiding loops that satisfies conformal restriction for simply connected domains. As we shall see later, it turns out that it satisfies also the general conformal restriction property.

In [15], [16] (see also Schramm's contribution in these proceedings), it is proved that the Hausdorff dimension of the outer boundary of a Brownian path is almost surely $4/3$ (the proof uses SLE considerations and we shall explain why later in this paper). Hence:

**Corollary 3.3.** *For the* (*up-to-constants*) *unique measure on fillings that satisfies conformal restriction, the boundary of the filling is almost surely a self-avoiding loop with dimension* $4/3$.

**3.2. The chordal case.** Suppose that for each $(D, A, B) \in \mathcal{T}$, we have the law $P_{D,A,B}$ of a random excursion from $A$ to $B$ in $D$. We say that the family $(P_{D,A,B})$ is conformally invariant if for any $D, A, B$ and any conformal map from $D$ onto some domain $D' = \Phi(D)$, the image measure of $P_{D,A,B}$ under $\Phi$ is the measure $P_{\Phi(D),\Phi(A),\Phi(B)}$.

This implies in particular that $P_{D,A,B}$ is invariant under any conformal map from $D$ onto itself that preserves the boundary points $A$ and $B$. For instance, for $D = \mathbb{H}$, $A = 0$ and $B = \infty$, this means that $P_{\mathbb{H},0,\infty}$ is scale-invariant (i.e. for each $\lambda > 0$, $\gamma$ and $\lambda\gamma$ have the same law modulo time-reparametrization). We then say that the probability measure $P_{D,A,B}$ is conformally invariant.

Conversely, if one has a probability measure $P$ on excursions from $A_0$ to $B_0$ in $D_0$ for some given triplet $(D_0, A_0, B_0)$ that is conformally invariant, one can simply define for each $D, A, B$ in $\mathcal{T}$ the measure $P_{D,A,B}$ to be the conformal image of $P$ under a conformal map from $(D_0, A_0, B_0)$ onto $(D, A, B)$. The obtained family $(P_{D,A,B})$ is then conformally invariant.

We say that the family $(P_{D,A,B})$ is restriction-invariant if for any $D, A, B$, and any simply connected subset $D'$ of $D$ such that the distance between $\{A, B\}$ and $D \setminus D'$ is positive (this implies in particular that $A$ and $B$ are on $\partial D'$), one has

$$P_{D,A,B}(\cdot \mid \gamma \subset D') = P_{D',A,B}(\cdot).$$

In other words, if $\gamma$ is defined under $P_{D,A,B}$, the conditional law of $\gamma$ given $\gamma \subset D'$ is exactly $P_{D',A,B}$.

**Definition 3.4.** We say that the probability measure $P_{D,A,B}$ for some $(D, A, B) \in \mathcal{T}$ satisfies conformal restriction if:

- It is conformally invariant.
- The conformally invariant family that it defines is restriction-invariant

Note that an excursion $\gamma$ from $A$ to $B$ in $D$ defines also a filling $K(\gamma)$, and that one can generalize the conformal restriction property to fillings also.

For a fixed triplet $D, A, B$, we call $\mathcal{D}_{D,A,B}$ the set of all simply connected domains $D' \subset D$ such that the distance between $D \setminus D'$ and $\{A, B\}$ is strictly positive. For each such $D'$, we define a conformal map from $D'$ back onto $D$ with $\Phi(A) = A$ and $\Phi(B) = B$. In the case where $\partial D$ is smooth in the neighborhood of $A$ and $B$, one can define $\Phi'(A)$ and $\Phi'(B)$ (which are real numbers) and note that the product of these two derivatives does not depend on which $\Phi$ (in the possible one-dimensional family of maps) one did choose. When $\partial D$ is not smooth in the neighborhood of $A$ and $B$, it is still possible to make sense of the quantity "$\Phi'(A)\Phi'(B)$" by conformal invariance (map $D$ onto the unit disc, and look at the corresponding quantity for the image of $A$, $B$ and $D'$). In short, the quantity $\Phi'(A)\Phi'(B)$ is a conformally invariant quantity that measures how smaller $D'$ is compared to $D$, seen from the two points/prime ends $A$ and $B$.

**Theorem 3.5** ([17]). *For each triple $(D, A, B) \in \mathcal{T}$, there exists exactly* (*and in particular: no more than*) *a one-parameter family of measures on fillings that satisfy conformal restriction. It is parametrized by a number $\alpha \in [5/8, \infty)$ and for each $\alpha$, the corresponding measure $P^{\alpha}_{D,A,B}$ is characterized by the property that for each $D' \in \mathcal{D}_{D,A,B}$,*

$$P^{\alpha}_{D,A,B}(K \subset D') = (\Phi'(A)\Phi'(B))^{\alpha}.$$

*Proof* (*sketch*). The uniqueness part is analogous to the loop case: By conformal invariance, we may choose $D$, $A$, $B$ to be $\mathbb{U}, -1, 1$. Then, the set $\mathcal{D} := \mathcal{D}_{\mathbb{U},-1,1}$ is the family of simply connected subsets $U$ of $\mathbb{U}$ such that $\mathbb{U} \setminus U$ is at positive distance from 1 and $-1$. For each such $U$, we define $\psi = \psi_U$ to be the unique conformal map from $U$ onto $\mathbb{U}$ such that $\psi(-1) = -1$, $\psi(1) = 1$ and $\psi'(-1) = 1$. The family of these conformal maps is closed under composition, and for two such maps $\psi_1$ and $\psi_2$, $(\psi_1 \circ \psi_2)'(1) = \psi_1'(1)\psi_2'(1)$.

Suppose that the measure $P$ on fillings of excursions from $-1$ to 1 in $\mathbb{U}$ satisfies conformal restriction. We then define for each such $U \in \mathcal{D}$,

$$A(\psi_U) = P(K \subset U).$$

Conformal restriction implies readily that $A(\psi_U \circ \psi_V) = A(\psi_U) \times A(\psi_V)$ for all $U$ and $V$ in $\mathcal{D}$, and this leads to the fact that there exists a positive constant $\alpha$ such that

$$P(K \subset U) = A(\psi_U) = \psi_U'(1)^{\alpha}.$$

But the probability measure $P$ is fully characterized by the knowledge of all the probabilities $P(K \subset U)$ for $U \in \mathcal{D}$.

It then remains to see that for each $\alpha \geq 5/8$, these identities indeed describe a probability measure on fillings, and that when $\alpha < 5/8$, no such measure exists. The way we prove this in [17] is that we explicitly construct the measure when $\alpha \geq 5/8$ using the Schramm–Loewner Evolution (SLE) process. For $\alpha < 5/8$, we also construct what would be the unique possible candidate (that satisfies a weaker condition – called the one-sided conformal restriction property – than the conformal restriction property that we described) for $P$ (via SLE or Brownian means), and we show that this candidate fails to satisfy the actual conformal restriction property. □

It is easy to check that the Brownian excursions from $A$ to $B$ in $D$ (and their fillings therefore also) defined by $e_{D,A,B}$ do satisfy conformal restriction for $\alpha = 1$, so that for $P^{1}_{D,A,B}$ the boundary of the filling is almost surely supported on sets of Hausdorff dimension $4/3$.

Let us give a partial description of the boundary of these fillings for general $\alpha$ in terms of Brownian excursions. Let us stick to case of the triplet $\mathbb{U}, -1, 1$. Suppose that $K$ is a filling satisfying conformal restriction. Then it turns out that $K \cap \partial\mathbb{U} = \{-1, 1\}$ and that the complement of $K$ in $\mathbb{U}$ consists of two connected components: The upper one $O^+$ such that $\partial O^+$ contains the upper half-circle $\partial_+ := \{e^{i\theta},\ \theta \in (0, \pi)\}$ and the lower one $O^-$, such that $\partial O^-$ contains the lower semi-circle $\partial_-$. The boundary

Figure 5. A Brownian excursion from $-1$ to 1 in the unit disc (sketch).

of $O^+$ (resp. $O^-$) then consists of the upper (resp. lower) semi-circle and a continuous curve $\gamma^+$ (resp. $\gamma^-$) joining $-1$ to $+1$ in $\mathbb{U}$. It is then not difficult to see that the law of $\gamma^+$ is characterized by the fact that for any $U \in \mathcal{D}$, such that $\mathbb{U} \setminus U$ is at positive distance of the lower semi-circle (i.e. $\mathbb{U} \setminus U$ is attached to the upper semi-circle)

$$P^\alpha(\gamma^+ \subset U) = \varphi'_U(1)^\alpha$$

(we will call $\mathcal{D}^+$ this subset of $\mathcal{D}$). One way to construct such a random curve uses a Poissonization argument and the Brownian excursion measure that we described earlier. Since a similar Poissonization argument will be useful in another setup a little bit later, let us briefly describe this classical idea in abstract terms:

Suppose that $N$ is a $\sigma$-finite measure without atoms on some space $\mathcal{X}$. We can define the law of a random countable family $X = \{X_j, j \in J\}$ of elements of $X$ in such a way that:

- For each $A_1, A_2 \subset \mathcal{X}$ in the $\sigma$-field on which $N$ is defined, such that $A_1 \cap A_2 = \emptyset$, the random families $X \cap A_1$ and $X \cap A_2$ are independent.
- For each $A_1$ as above, the probability that $X \cap A_1$ is empty equals $\exp(-N(A_1))$.

The law of $X$ is in fact characterized by these two properties. It is easy to see that for each $A$, the cardinality of $X \cap A$ is a Poisson random variable with mean $N(A)$ (so that it is a.s. infinite if and only if $N(A) = \infty$). $X$ is called a Poisson point process with intensity $N$.

Note that if $X_1$ and $X_2$ are two independent Poisson point processes on the same space $\mathcal{X}$ with respective intensity $N_1$ and $N_2$, then $X_1 \cup X_2$ is a Poisson point process with intensity $N_1 + N_2$.

Using this idea, one can define on the same probability space a collection $(X_c, c \geq 0)$ of Poisson point processes in such a way that $X_c \subset X_{c'}$ for all $c \leq c'$, and such that the intensity of $X_c$ is $cN$. One intuitive way to view this is to say that with time, elements of $\mathcal{X}$ appear independently. During a time-interval $dt$, an element of

a set $A \subset \mathcal{X}$ will appear with probability $dt \times N(A)$. Then, $X_c$ denotes the family of elements that did appear before time $c$.

Let us now use this construction for a measure $N$ on the space of excursions in $\mathbb{U}$. More precisely, we define $\mathcal{X}$ the set of excursions in $\mathbb{U}$ that start and end on the lower semi-circle $\partial_-$, and we define $N$ to be $\mathrm{exc}_{\mathbb{U}}$ restricted to this set of excursions.

Hence, for each $c$, the previous procedure defines a random countable collection of Brownian excursions $E_c = (e_j, j \in J_c)$ starting and ending on the negative half-circle. Despite the fact that this collection is almost surely infinite (because the total mass of $N$ is infinite), the total number of excursions of diameter greater than $\varepsilon$ is almost surely finite for all positive $\varepsilon$ (because the $N$-mass of this set of excursions is finite). In particular, this implies that the "upper boundary" $\gamma^+$ of the union of all excursions in $E_c$ does not intersect the upper semi-circle $\partial_+$. It does not exit a given $U \in \mathcal{D}^+$ if and only if no excursion in $E_c$ does exit $U$, and by definition, this happens with probability $\exp(-cN(\{\gamma : \gamma \not\subset U\}))$.

The conformal restriction property of the excursion measure shows that for each $U \in \mathcal{D}^+$, the image under $\varphi_U$ of the measure $N$ restricted to the set of excursions that stay in $U$ is exactly equal to $N$. It follows readily from this fact that $\exp(-N(\{\gamma : \gamma \not\subset U\})) = \varphi'_U(1)^{\alpha_1}$ for some $\alpha_1$. Hence, for each $\alpha > 0$, if one chooses $c = \alpha/\alpha_1$, the curve $\gamma^+$ does indeed satisfy $P(\gamma^+ \subset U) = \varphi'_U(1)^{c\alpha_1} = \varphi'_U(1)^{\alpha}$.

The fact that $\alpha < 5/8$ is not possible corresponds to the fact that the probability that $\gamma^+$ goes "below" the origin becomes larger than $1/2$, which is not possible for symmetry reasons if it is equal to the upper boundary of a filling satisfying conformal restriction.

For more precise statements and also other possible descriptions of the joint law of $(\gamma^+, \gamma^-)$, see [17], [39], [40].

## 4. Related models

So far, we have defined only measures on Brownian curves, and we have basically shown that any measure satisfying conformal restriction defines the same outer boundary as that of these Brownian measures. The theory becomes interesting when we note that some a priori different measures do also satisfy conformal restriction.

**4.1. Percolation.** We now very briefly describe the percolation model that has been proved by Smirnov [35] to be conformally invariant in the scaling limit. Consider the honeycomb lattice (the regular tiling of the plane by hexagons) with mesh size $\delta$. Each hexagon is colored independently in black or in white with probability $1/2$. Then, we are interested in the connectivity properties of the set of white (resp. black) cells. We call white (resp. black) cluster a connected component of the union of the white (resp. black) cells. This model is sometimes called "critical site-percolation on the triangular lattice".

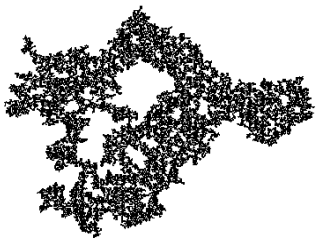


Figure 6. A rescaled large percolation cluster.

By now classical arguments due to Russo, Seymour and Welsh show that the number of clusters that are of diameter $\varepsilon > 0$ in the unit disc remains tight when $\delta \to 0$. This suggests the existence of a scaling limit for the joint law on all clusters when $\delta$ (in an appropriately chosen topology). Smirnov [35] proved the existence of the limit of certain observables (the crossing probabilities) and their conformal invariance.

A consequence of this result is [36] that it is possible to use SLE computations from [14], [15] and earlier results from Kesten [11] to deduce the existence and the value of the critical exponents for critical percolation as predicted by theoretical physicists such as Cardy, Duplantier, Saleur (see e.g. the references in [36]). But, we would here like to focus on the conformal restriction aspect of the scaling limit of percolation and its consequences. We will remain on a heuristic level, but what follows can be made rigorous:

A percolation configuration is described by its white (say) clusters $(C_j,\ j \in J)$. Smirnov's result can be shown to imply (see [5]) the convergence in law of this family when $\delta \to 0$ to the joint law of a collection of "clusters" $(C_j,\ j \in J)$ in the plane. A slightly weaker statement is that the measure on clusters $\pi^\delta$ that assigns to each possible cluster the probability that this cluster indeed occurs converges when $\delta \to 0$ towards a measure on "clusters" $\pi$. The measure $\pi$ satisfies conformal restriction. This is due to the combination of conformal invariance (due to Smirnov's result) and of the independence properties of percolation from which restriction immediately follows in the scaling limit. Hence:

**Proposition 4.1.** *The measure $\pi$ on scaling limits of critical percolation clusters satisfies conformal restriction.*

So, $\pi$ defines exactly the same fillings as (a multiple) of the Brownian loop measure $M$, and it defines a measure on outer boundaries that is exactly a multiple of $\mu$. In other words, the shape of the outer perimeter of a very large percolation cluster has (in the scaling limit) the same law than the outer boundary of a Brownian loop.

**4.2. The self-avoiding walk conjectures.** A classical open problem is to understand the behavior of very long self-avoiding paths, sampled uniformly among all such long self-avoiding paths on some planar lattice with a given starting point and a given length $N$, in the limit when $N \to \infty$.

It is believed that in the scaling limit (for regular periodic lattices with some rotational symmetry) these paths exhibit conformal invariance properties. This led to various striking predictions by theoretical physicists concerning this model and its critical exponents.

For instance, it is believed that the diameter of a typical self-avoiding path with $N$ steps is of the order of $N^{3/4}$. This can be loosely phrased in terms of "fractal dimension" since it means that one requires $N$ steps of size $N^{-3/4}$ to cover a long self-avoiding walk of macroscopic size on the lattice $N^{-3/4}\mathbb{Z}^2$. More precisely, this could mean that in the scaling limit, self-avoiding walks converge in law to some continuous measure on paths supported on the set of paths with dimension equal to $4/3$.

Note that the number of self-avoiding walks of length $N$ on $\mathbb{Z}^2$ that start at the origin can easily (via a sub-multiplicativity argument) be shown to behave like $\lambda^{N+o(N)}$ when $N \to \infty$, where $\lambda$ is a positive real number called the connectivity constant of $\mathbb{Z}^2$. One of the striking conjectures in this field is the more precise prediction $\lambda^N N^{11/32+o(1)}$ by Nienhuis [28].

Here are two possible ways to state this existence of scaling-limit conjecture (in the case of the square lattice):

- Self-avoiding loops: The measure on self-avoiding loops on $\delta\mathbb{Z}^2$ that assigns a mass $\lambda^{-N}$ to each loop with $N$ steps has a (non-trivial) limit when $\delta \to 0$.
- Excursions: The probability measure on self-avoiding excursions from $-1$ to $1$ in an approximation of $\mathbb{U}$ by a sublattice of $\delta\mathbb{Z}^2$ that assigns a probability proportional to $\lambda^{-N}$ to each excursion with $N$ steps converges (when $\delta \to 0$) to a (non-trivial) scaling limit.

In the first case, the scaling limit is then a measure $S$ supported on the set of loops in the plane. In the second one, it is then a probability measure $P^S$ on the set of excursions from $-1$ to $1$ in $\mathbb{U}$.

If one assumes furthermore that these measures exhibit conformal invariance properties, then $S$ should be a measure on self-avoiding loops satisfying conformal restriction: By the previously described results, it is therefore a multiple of the measure $\mu$ on outer boundaries of Brownian loops and of the measure on outer boundaries of percolation clusters. Similarly, we get that $P^S$ should satisfy chordal conformal restriction. Hence, it should be a measure on excursions without double points that coincides with one of the $P^\alpha$'s. This gives an explanation (but not a proof) of the $4/3$-dimension conjecture for self-avoiding walks.

Let us note that in his book [26], Mandelbrot had already proposed the name "self-avoiding Brownian motion" for the outer boundary of a planar Brownian loop.

The above results show that this would be indeed an appropriate name.

## 5. Related SLEs

The (chordal) Schramm–Loewner Evolutions (SLE) first introduced in [31] are conformally invariant random planar excursions in a domain with prescribed endpoints. They are defined via iterations of random conformal maps and they are the only ones satisfying a certain Markov property. Since the discrete analogue of this Markovian property is obviously satisfied by the interfaces of many discrete lattice-models from statistical physics (including for instance percolation), this shows that if these discrete interfaces converge to conformally invariant scaling limits, then they have to be one of the SLE curves. For details on the definition and properties of SLEs, their relations (conjectured and proved) to lattice-models, there are now many surveys, lecture notes, a book (e.g. [13], [37] and the references therein); see also Schramm's contribution to the present ICM proceedings.

There exists a one-parameter family of SLE's: For each $\kappa > 0$, the SLE with parameter $\kappa$ (in short: $\mathrm{SLE}_\kappa$) is a mathematically well-defined random planar excursion joining prescribed boundary points in a simply connected domain [30], [18]. One can then see if these random excursions satisfy conformal restriction (in the chordal case). It turns out that:

**Proposition 5.1** ([17]). *$SLE_{8/3}$ is a random excursion without double points that satisfies chordal restriction. Its law is exactly $P^{5/8}$. No other SLE satisfies chordal conformal restriction.*

In fact, one can prove that it is the only measure supported on excursions without double points that satisfies chordal conformal restriction (i.e. that for all $\alpha > 5/8$, the measure $P^\alpha$ is not supported on self-avoiding curves). Hence, the $\mathrm{SLE}_{8/3}$ is the conjectural scaling limit of self-avoiding excursions, i.e. $P^S = P^{5/8}$. Not surprisingly given all what we have said so far, it can be proved directly that it is supported on the set of excursions with Hausdorff dimension $4/3$ [17], [2]. The computation of the critical exponents for SLE (e.g. [14], [15]) allow also to recover the physicists' predictions on critical exponents such as the $11/32$ mentioned above (see e.g. [19]).

Also, there is a rather direct relation between discrete self-avoiding loops and self-avoiding excursions (the self-avoiding excursion tells how to finish a loop if we know part of it). This suggests a direct relation between the outer boundaries of planar Brownian loops and the $\mathrm{SLE}_{8/3}$ processes. Indeed (see e.g. [23]), it is possible to define a measure on $\mathrm{SLE}_{8/3}$ loops and to see that it is a measure on self-avoiding loops in the plane that satisfies conformal restriction:

**Proposition 5.2.** *The measure $\mu$ can be viewed as a measure on $SLE_{8/3}$ loops.*

In fact, this has a deeper consequence, which is not really surprising if one thinks

of $\mu$ in terms of the conjectural scaling limit $S$ of the measure on discrete self-avoiding loops:

**Theorem 5.3** ([42]). *The measure $\mu$ on self-avoiding loops satisfies conformal restriction also for non-simply connected domains $D$.*

A particular instance of the theorem is that the measure $\mu$ is invariant under the inversion $z \mapsto 1/z$. This implies [42] that the inner boundaries of Brownian loops (and those of the scaling limits of critical percolation clusters) have exactly the same distribution than the outer boundaries. More precisely, if one looks at the boundary of the connected component that contains the origin of the complement of a Brownian loop (defined under $M$) then it is defined under exactly the same measure as the outer boundary. This is by no means an obvious fact.

Another consequence is the following:

**Corollary 5.4** ([42]). *It is possible to extend the definition of the planar measure $\mu$ on self-avoiding loops to any Riemann surface (possibly with boundaries) in such a way that conformal restriction still holds.*

This gives a direct description of various conformally invariant quantities in the framework of Riemann surfaces.

The $SLE_6$ process can be shown (see e.g. [14], [15]) to be the only SLE satisfying a so-called locality property that makes it the only possible candidate for the (conformally invariant) scaling limit of percolation interfaces. In fact, using Smirnov's result [35] and ideas, it is possible to deduce [5] that $SLE_6$ is indeed this scaling limit for critical percolation on the triangular lattice. Hence, it should not be surprising that it is possible to define directly (from the definition of $SLE_6$) conformally invariant measures on loops and excursions that satisfy conformal restriction (see e.g. [17]). This is one of the ways to see that chordal restriction for $\alpha = 2$ is very closely related to the loop measure $\mu$.

## 6. Restriction defect

Most models arising from statistical physics should however not satisfy conformal restriction in the scaling limit. Self-avoiding walks and percolation are in this respect rather exceptional cases. We now describe how one can extend the conformal restriction property to cover the more generic cases. It is useful to start with a specific model to illustrate the basic ideas and to show why the Brownian loop-soup can be useful.

**6.1. Loop-erased random walks.** Suppose that $S = (S_n, n \leq N)$ is a discrete nearest neighbor-walk of length $N$ on a finite connected graph $G$. It is as a path joining the two points $o = S_0$ and $e = S_N$ that can have double points. One can however associate to $S$ a path from $o$ to $e$ without double-points by following $S$ and

erasing the loops as they appear. This gives rise to the loop-erasure $L = L(S)$ of $S$. It is the only simple path from $o = L_0$ to $e = L_p$ (the length $p$ of $L$ is not greater than $N$ but it can be smaller and it depends on the length of the loops erased during this procedure) with the property that for each $i \leq p-1$, $L_{i+1} = S_{n_i+1}$, where $n_i = \sup\{n \leq N : S_n = L_i\}$.

If we are given the two points $o$ and $e$, we can choose $S$ randomly to be a simple random walk on the graph, started at $o$ and stopped at its first hitting of $e$. Its loop-erasure $L = L(S)$ is then the so-called loop-erased random walk from $o$ to $e$. It has many nice combinatorial features, that are not obvious at first sight. For instance, the law of the loop-erased random walk from $o$ to $e$ and of the loop-erased random walk from $e$ to $o$ are the same (modulo time-reversal of course). It can also be interpreted as the law of the unique (simple) path joining $o$ to $e$ in a spanning tree chosen uniformly among all spanning trees of the graph $G$ (i.e. choose uniformly a subgraph of $G$ with just one connected component but no cycle, and look at the unique path joining $o$ to $e$ in this subgraph). This result by Pemantle [29] has been extended by Wilson into a complete construction of a uniformly chosen spanning tree of $G$ using loop-erased random walks [43]. It shows that loop-erased random walks belong to a wider general class of models from statistical physics (the random-cluster models) that includes also the Ising models.

A fine-grid approximation of the Brownian excursion measure $e_{\mathbb{U},-1,1}$ goes as follows: Consider a fine-mesh approximation of the unit disc with two boundary points $o$ and $e$ close to $-1$ and 1, and consider a simple random walk started from $o$, stopped at $e$, and conditioned to exit $\mathbb{U}$ through $e$.

**Theorem 6.1** ([18]). *The loop-erasure of this discrete excursion converges when the mesh-size converges to zero to a conformally invariant scaling limit, the $SLE_2$ from $-1$ to 1 in $\mathbb{U}$. Similarly, for any triplet $(D, A, B) \in \mathcal{T}$, the loop-erasure of a fine-grid approximation of an excursion defined under $e_{D,A,B}$ converges to the $SLE_2$ from $A$ to $B$ in $D$.*

For a given $U \subset \mathbb{U}$ that still has $-1$ and 1 on its boundary, it happens with positive probability that the loop-erasure $L$ of the discrete excursion $S$ stays in $U$, but that the path $S$ does exit $U$ (i.e. one of the erased loops went out of $U$). This feature pertains in the scaling limit and shows that conformal restriction is not satisfied by $SLE_2$. The lack of restriction can be quantified in terms of the erased random walk loops (i.e. in the scaling limit in terms of a quantity involving Brownian loops). More precisely, for a given simple nearest neighbor path from $o \sim -1$ to $e \sim 1$ on the $\delta\mathbb{Z}^2$-approximation of $U \subset \mathbb{U}$, the ratio between the probability that $L = l$ for the LERW from $o$ to $e$ in $U$ and the probability that $L = l$ for the LERW in $\mathbb{U}$ is given by

$$F_\delta(l) = \mathrm{cst}(U)\, P_{\mathbb{U}}(\text{none of the erased loops did exit } U \mid L = l).$$

This function $F_\delta$ converges to a non-trivial function $F$ when $\delta \to 0$ that measures the restriction-defect of $SLE_2$ and that can be expressed in terms of Brownian loops.

**6.2. The Brownian loop soup.** Consider the (properly normalized) Brownian loop-measure $M$. Recall that it is a measure on the set of unrooted Brownian loops in the entire plane. For each $c > 0$, we define a Poisson point process with intensity $cM$. This is a random countable collection $\{b_j, j \in J\}$ of Brownian loops in the plane.

For each domain $D$, we define $J(D) = \{j \in J : b_j \subset D\}$. It is clear from the definition that this corresponds to a Poisson point process with intensity $cM_D$.

In [23], we show that:

**Proposition 6.2** ([23]). *The function $F(l)$ is equal to the probability that no loop in the loop-soup with intensity $2M_{\mathbb{U}}$ intersects both the excursion $l$ and the complement of $U$.*

This indicates that the loops that have been erased correspond to the loops in the loop-soup that the path $l$ intersects. This is not so surprising if one thinks of Wilson's algorithm (that in some sense shows that the law of the constructed uniform spanning tree is independent of the erased loops).

It shows [23] that if one adds to an $\mathrm{SLE}_2$ the loops that it intersects in a Brownian loop-soup, one recovers exactly a path satisfying conformal restriction (in fact with parameter $\alpha = 1$, the one of the Brownian excursion $\mathrm{exc}_{\mathbb{U},-1,1}$).

A similar coupling of the $\mathrm{SLE}_\kappa$'s for $\kappa < 8/3$ with a Brownian loop-soup of parameter $c = c(\kappa) = (8 - 3\kappa)(6 - \kappa)/2\kappa$. By adding the loops of this loop-soup to the SLE curve, one compensates its lack of restriction and constructs a filling that satisfies conformal restriction with parameter $\alpha = (6 - \kappa)/2\kappa$. These relations correspond to the relation between the central charge $(-c)$, the highest weight $(\alpha)$ and the degeneracy factor $(\kappa/4)$ of degenerate highest-weight representations of the Virasoro Algebra, as predicted by conformal field theory (see e.g. [9], [1], [3]).

**6.3. Loop-soup clusters, CLEs.** This does not describe the type of restriction-defects of the SLE's with parameter $\kappa > 8/3$ that should arise as scaling limits of various lattice models, corresponding in the physics language to models with positive central charge. Loosely speaking, these are the curves that are attracted by the boundaries of a domain (as opposed for instance to the $\mathrm{SLE}_2$ that was "repelled" from the boundary). The previous case $\kappa < 8/3$ corresponded to a negative central charge.

For this, it is useful to consider the geometry of the union of all loops in a loop-soup of intensity $c\mu_{\mathbb{U}}$ (recall that the measure $\mu_{\mathbb{U}}$ corresponds to the outer boundaries of the Brownian loops defined by $M_{\mathbb{U}}$). This loop-soup is a countable collection $C_c = \{\ell_j, j \in J_c\}$ of self-avoiding loops in the unit disc that can overlap with each other. Recall that can couple all $C_c$'s in such a way that $c \mapsto C_c$ is increasing.

When $c$ is large and fixed, it is not difficult to see that almost surely every point in $\mathbb{U}$ is surrounded by a loop in $C_c$, so that all the loops hook up into one single connected component i.e. the set $\bigcup_{j \in J_c} \ell_j$ has just one connected component.

On the other hand, when $c$ is small, it is also easy for instance by coupling this problem with the so-called fractal percolation (sometimes also called Mandelbrot percolation) studied in [7], [27] to see that this phenomenon does not pertain: The

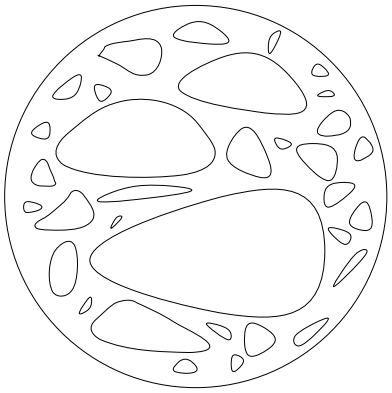


Figure 7. A CLE (very very sketchy).

set $\bigcup_{j\in J_c} \ell_j$ has countably many connected components. The outermost boundaries of these clusters of loops define a family of non-overlapping and non-nested loops $\boldsymbol{u}^c = \{u_i^c,\ i \in I_c\}$ in $\mathbb{U}$.

This leads to the following definition [34]:

**Definition 6.3.** Suppose that $\boldsymbol{u} = \{u_i, i \in I\}$ is a random collection of non-intersecting and non-nested self-avoiding loops in $\mathbb{U}$. We say that it is a simple conformal loop-ensemble (CLE) if the following properties hold:

- It is invariant under the conformal transformation from $\mathbb{U}$ onto itself. This allows to define the law $P_U$ of the collection of loops in any simply connected domain $U$ by taking the conformal image of $\boldsymbol{u}$.
- Let $U$ be any simply connected subset of $\mathbb{U}$ with $d(\mathbb{U} \setminus U, 1) > 0$. Consider $I' = \{i \in I, u_i \not\subset U\}$ and let $\tilde{U}$ denote the connected component of $U \setminus \bigcup_{i\in I'} u_i$ that has 1 on its boundary. Then, conditionally on $\{u_i, i \in I'\}$, the law of $\{u_i, i \in I \text{ and } u_i \subset \tilde{U}\}$ is $P_{\tilde{U}}$.

Loosely speaking, this means that each loop (once it is discovered) plays the role of the boundary of the domain in which the others are yet to be discovered. Note that a CLE almost surely is an infinite collection of loops (because the number of loops contained in $\mathbb{U}$ and in $\tilde{U} \subset \mathbb{U}$ have the same law).

The previous considerations show that the outermost boundaries of cluster of loops for sub-critical (i.e. for small $c$) loop-soups, are conformal loop ensembles (so that CLEs exist). This gives rise to measures on loops that do not satisfy conformal restriction, but have the same type of restriction defect as that of SLEs for $\kappa \in (8/3, 4]$. The intensity $c$ of the loop-soup corresponds to the central charge of the model.

Conformal loop-ensembles (and SLEs) arise also in the context of level-lines (or flow-lines) of the Gaussian Free Field [33] in the ongoing work of Oded Schramm and

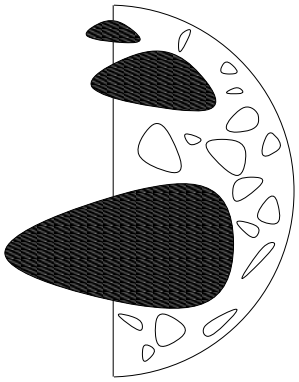


Figure 8. The CLE in $\tilde{U}$ when $U$ is the semi-disc.

Scott Sheffield [32]. Combining all these arguments should [34] describe all CLEs as loop-cluster boundaries and their boundaries as SLE loops for $\kappa \leq 4$.

**Acknowledgments.** I would like to express many thanks to Greg Lawler and Oded Schramm for the opportunity to interact and work with them during these last years. I also thank Pierre Nolin for Figures 4 and 6.

## References

[1] Bauer, M., Bernard, D., Conformal transformations and the SLE partition function martingale. *Ann. Henri Poincaré* **5** (2004), 289–326.

[2] Beffara, V., The dimensions of SLE curves. math.PR/0211322, 2002.

[3] Belavin, A. A., Polyakov, A. M., Zamolodchikov, A. B., Infinite conformal symmetry in two-dimensional quantum field theory. *Nuclear Phys. B* **241** (1984), 333–380.

[4] Burdzy, K., Lawler, G. F., Non-intersection exponents for random walk and Brownian motion II. Estimates and application to a random fractal. *Ann. Probab.* **18** (1990), 981–1009.

[5] Camia, F., Newman, C., The Full Scaling Limit of Two-Dimensional Critical Percolation. Preprint, 2005; arXiv:math.PR/0504036.

[6] Cardy, J. L., Conformal invariance and surface critical behavior. *Nuclear Phys. B* **240** (1984), 514–532.

[7] Chayes, J. T., Chayes, L., Durrett, R., Connectivity properties of Mandelbrot's percolation process. *Probab. Theory Related Fields* **77** (1988), 307–324.

[8] Duplantier, B., Conformal fractal geometry and boundary quantum gravity. In *Fractal Geometry and applications, a jubilee of Benoît Mandelbrot*. Proc. Symp. Pure Math. **72**, Part II, Amer. Math. Soc., Providence, RI, 2004, 365–482.

[9] Friedrich, R., Werner, W., Conformal restriction, highest-weight representations and SLE. *Comm. Math. Phys.* **243** (2003), 105–122.

[10] Garban, C., Trujillo-Ferreras, J. A., The expected area of the Brownian loop is $\pi/5$. *Comm. Math. Phys.* **264** (3) (2006), 797–810.

[11] Kesten, H., Scaling relations for 2D-percolation. *Comm. Math. Phys.* **109** (1987), 109–156.

[12] Kozdron, M., On the scaling limit of simple random walk excursion measure in the plane. Preprint, 2005; arXiv:math.PR/0506337.

[13] Lawler, G. F., *Conformally invariant processes in the plane*. Math. Surveys Monogr. 144, Amer. Math. Soc., Providence, RI, 2005.

[14] Lawler, G. F., Schramm, O., Werner, W., Values of Brownian intersection exponents I: Half-plane exponents. *Acta Math.* **187** (2001), 236–273.

[15] Lawler, G. F., Schramm, O., Werner, W., Values of Brownian intersection exponents II: Plane exponents. *Acta Math.* **187** (2001), 275–308.

[16] Lawler, G. F., Schramm, O., Werner, W., The dimension of the Brownian frontier is $4/3$. *Math. Res. Lett.* **8** (2001), 401–411.

[17] Lawler, G. F., Schramm, O., Werner, W., Conformal restriction properties. The chordal case. *J. Amer. Math. Soc.* **16** (2003), 917–955.

[18] Lawler, G. F., Schramm, O., Werner, W., Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab.* **32** (2004), 939–996.

[19] Lawler, G. F., Schramm, O., Werner, W., On the scaling limit of planar self-avoiding walks. In *Fractal Geometry and applications, a jubilee of Benoît Mandelbrot*, Proc. Symp. Pure Math. 72, Part II, Amer. Math. Soc., Providence, RI, 2004, 339–364.

[20] Lawler, G. F., Trujillo-Ferreras, J. A., Random walk loop-soup. *Trans. Amer. Math. Soc.*, to appear.

[21] Lawler, G. F., Werner, W., Intersection exponents for planar Brownian motion. *Ann. Probab.* **27** (1999), 1601–1642.

[22] Lawler, G. F., Werner, W., Universality for conformally invariant intersection exponents. *J. Europ. Math. Soc.* **2** (2000), 291–328.

[23] Lawler, G. F., Werner, W., The Brownian loop-soup. *Probab. Theory Related Fields* **128** (2004), 565–588.

[24] Le Gall, J. F., Some properties of planar Brownian motion. In *École d'Été de Probabilités de Saint-Flour XX—1990* (ed. by P. L. Hennequin), Lecture Notes in Math. 1527, Springer-Verlag, Berlin 1992, 111–235.

[25] Lévy, P., *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris 1948.

[26] Mandelbrot, B. B., *The Fractal Geometry of Nature*. W. H. Freeman and Co., San Francisco, CA, 1982.

[27] Meester, R., Roy, R., *Continuum Percolation*. Cambridge Tracts in Math. 119, Cambridge University Press, Cambridge 1996.

[28] Nienhuis, B., Exact critical exponents for the $O(n)$ models in two dimensions. *Phys. Rev. Lett.* **49** (1982), 1062–1065.

[29] Pemantle, R., Choosing a spanning tree for the integer lattice uniformly. *Ann. Probab.* **19** (1991), 1559–1574.

[30] Rohde, S., Schramm, O., Basic properties of SLE. *Ann. of Math.* (2) **161** (2005), 879–920.

[31] Schramm, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118** (2000), 221–288.

[32] Schramm, O., Sheffield, S., in preparation.

[33] Sheffield, S., Gaussian Free Fields for mathematicians. Preprint, 2003; arXiv:math.PR/0312099.

[34] Sheffield, S., Werner, W., in preparation.

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

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

[37] Werner, W., Random planar curves and Schramm-Loewner Evolutions. In *Lectures on probability theory and statistics*, Lecture Notes in Math. 1840, Springer-Verlag, Berlin 2004, 107–195.

[38] Werner, W., SLEs as boundaries of clusters of Brownian loops. *C. R. Acad. Sci. Paris Sér. I Math.* **337** (2003), 481–486.

[39] Werner, W., Girsanov's Theorem for SLE($\kappa$, $\rho$) processes, intersection exponents and hiding exponents. *Ann. Fac. Sci. Toulouse* **13** (2004), 121–147.

[40] Werner, W., Conformal restriction and related questions. *Probab. Surv.* **2** (2005), 145–190.

[41] Werner, W., Some recent aspects of conformally invariant systems. Lecture Notes from Les Houches summer school, Preprint, 2005; arXiv:math.PR/0511268.

[42] Werner, W., The conformal invariant measure on self-avoiding loops. Preprint, 2005; arXiv:math.PR/0511605.

[43] Wilson, D. B., Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing*, ACM, New York 1996, 296–303.

Université Paris-Sud, Laboratoire de mathématiques, Bât. 425, 91405 Orsay, France
and
Ecole Normale Supérieure, Département de mathématiques et applications, 45, rue d'Ulm, 75230 Paris cedex 05, France

E-mail: wendelin.werner@math.u-psud.fr

# The complexity of generating functions for integer points in polyhedra and beyond

Alexander Barvinok*

**Abstract.** Motivated by the formula for the sum of the geometric series, we consider various classes of sets $S \subset \mathbb{Z}^d$ of integer points for which an a priori "long" Laurent series or polynomial $\sum_{m \in S} \boldsymbol{x}^m$ can be written as a "short" rational function $f(S; \boldsymbol{x})$. Examples include the sets of integer points in rational polyhedra, integer semigroups, and Hilbert bases of rational cones, among others. We discuss applications to efficient counting and optimization and open questions.

**Mathematics Subject Classification (2000).** Primary 05A15; Secondary 68W30, 11P21, 52C07, 11H06.

**Keywords.** Lattice point, rational polytope, generating function, rational function, Laurent polynomial, integer semigroup, Hilbert basis, efficient counting, computational complexity.

## 1. Introduction

Our inspiration comes from a formula for the sum of a finite geometric series:

$$\sum_{m=0}^{n} x^m = \frac{1 - x^{n+1}}{1 - x}. \tag{1.1}$$

We look at the formula from several points of view.

Geometrically, the left hand side of (1.1) represents the sum over all integer points in a one-dimensional polytope. Namely, with every integer point $m$ we associate a monomial $x^m$ and then consider the sum over all integer points in the interval $[0, n]$.

From the computational complexity point of view, the left hand side of (1.1) is a "long" polynomial whereas the right hand side of (1.1) is a "short" rational function. More precisely, to write an integer $m$ we need about $\log m$ digits or bits. Consequently, to write the left hand side of (1.1), we need about $n \log n$ bits. On the other hand, to write the right hand side of (1.1) we need only about $\log n$ bits. Thus the left hand side is exponentially longer than the right hand side.

Finally, let us read (1.1) from right to left. We can ask how to extract various facts about the set $S$ of integer points in the interval $[0, n]$ from the rational function

*The author is grateful to Microsoft (Redmond) for hospitality during his work on this paper. This work was partially supported by NSF Grant DMS 0400617

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

encoding. For example, to compute the number $|S|$ of points we substitute $x = 1$ into the right hand side of (1.1). Although $x = 1$ is a pole of the rational function, we can compute the desired value by applying l'Hospital's rule.

Let $\mathbb{R}^d$ be Euclidean space with the standard basis $e_1, \ldots, e_d$, so a point $x \in \mathbb{R}^d$ is identified with the $d$-tuple $x = (\xi_1, \ldots, \xi_d)$ of its coordinates, and let $\mathbb{Z}^d \subset \mathbb{R}^d$ be the standard integer lattice, that is the set of points with integer coordinates. With every integer point $m = (\mu_1, \ldots, \mu_d)$ we associate the Laurent monomial

$$\boldsymbol{x}^m = x_1^{\mu_1} \ldots x_d^{\mu_d}$$

in $d$ complex variables $\boldsymbol{x} = (x_1, \ldots, x_d)$. We agree that $x_i^0 = 1$.

Let $S \subset \mathbb{Z}^d$ be a finite set and let us consider the sum

$$f(S; \boldsymbol{x}) = \sum_{m \in S} \boldsymbol{x}^m.$$

Thus $f(S; \boldsymbol{x})$ is a Laurent polynomial that is the generating function of the set $S$. We are interested in the following general questions:

• For which sets $S \subset \mathbb{Z}^d$ a potentially long Laurent polynomial $f(S; \boldsymbol{x})$ can be written as a short rational function?

• What information about the set $S$ can be extracted from $f(S; \boldsymbol{x})$ given as a short rational function?

The paper is organized as follows.

In Section 2, we discuss necessary preliminaries from the theory of computational complexity, define what "long" and "short" means and show that if $S$ is the set of integer points in a rational polyhedron $P \subset \mathbb{R}^d$ then the generating function $f(S; \boldsymbol{x})$ can be computed in polynomial time as a short rational function, provided the dimension $d$ of the ambient space is fixed in advance. We discuss applications to efficient counting and optimization and practical implementations of the algorithms.

In Section 3, we discuss what information can we extract from a set $S \subset \mathbb{Z}^d$ *defined* by its generating function $f(S; \boldsymbol{x})$ written as a rational function. In particular, we show that if $S_1, S_2 \subset \mathbb{Z}^d$ are two finite sets defined by their rational generating functions $f(S_1; \boldsymbol{x})$ and $f(S_2; \boldsymbol{x})$, then the generating function $f(S; \boldsymbol{x})$ of their intersection $S = S_1 \cap S_2$ can be computed in polynomial time as a rational function.

In Section 4, we show that if $S \subset \mathbb{Z}_+$ is an integer semigroup with a fixed number $d$ of generators, then $f(S; x)$ can be computed in polynomial time as a short rational function. This result is obtained as a corollary of a more general result that the *projection* of the set of integer points in a rational polytope admits a polynomial time computable rational generating function. We mention some other examples such as Hilbert bases of rational cones.

In Section 5, we consider the results of Sections 2 and 4 in the general context of Presburger arithmetic. We argue that the "natural" class of sets $S \subset \mathbb{Z}^d$ with short rational generating functions $f(S; \boldsymbol{x})$ would have been the class of sets defined by

formulas of Presburger arithmetic where all combinatorial parameters (the number of variables and Boolean operations) are fixed and only numerical constants are allowed to vary. As the paper is being written, this is still a conjecture.

In Section 6, we try to identify the natural boundaries of the developed theory. We also discuss the emerging picture of what happens if the dimension $d$ of the ambient space is allowed to grow.

## 2. Rational polyhedra

Formula (1.1) admits an extension to general rational polyhedra.

**Definition 2.1.** The set $P \subset \mathbb{R}^d$ of solutions to a system of finitely many linear inequalities is called a *polyhedron*:

$$P = \left\{(\xi_1, \ldots, \xi_d) : \sum_{j=1}^{d} \alpha_{ij}\xi_j \le \beta_i,\ i = 1, \ldots, n\right\}. \tag{2.1}$$

Here $\alpha_{ij}$ and $\beta_i$ are real numbers. A bounded polyhedron is called a *polytope*. A polyhedron $P$ is called *rational* if in (2.1) one can choose all $\alpha_{ij}$ and $\beta_i$ integer.

To state an analogue of formula (1.1) we need to discuss the notion of the input size. As we remarked earlier, to write an integer $a$ we need roughly $\lceil \log_2(|a|+1)\rceil + 1$ bits. Consequently, to define a rational polyhedron $P \subset \mathbb{R}^d$ by the inequalities (2.1) we need about

$$\mathcal{L} = n(d+1) + \sum_{i,j} \lceil \log_2(|\alpha_{ij}| + 1)\rceil + \sum_{i} \lceil \log_2(|\beta_i| + 1)\rceil \tag{2.2}$$

bits. The number $\mathcal{L}$ is called the *input size* of representation (2.1) of $P$.

We are interested in the computational complexity of formulas and algorithms. In particular, we are interested in *polynomial time* algorithms, that is, in the algorithms whose running time is at most $\mathcal{L}^{O(1)}$, where $\mathcal{L}$ is the input size. In what follows, often the dimension $d$ of the ambient space will be fixed in advance and the algorithms will run in polynomial time *for any fixed dimension* $d$. In other words, the running time of such an algorithm is at most $\mathcal{L}^{\phi(d)}$ for some function $\phi$. We use [28] as a general reference in the area of computational complexity and algorithms.

Let $P \subset \mathbb{R}^d$ be a rational polyhedron with a vertex (equivalently, a non-empty polyhedron without lines), possibly unbounded, and let $S = P \cap \mathbb{Z}^d$ be the set of integer points in $P$.

To simplify notation, we denote the generating function

$$f(S; \boldsymbol{x}) = \sum_{m \in S} \boldsymbol{x}^m,$$

where $S = P \cap \mathbb{Z}^d$, just by $f(P, \boldsymbol{x})$.

It is not hard to show that there exists a non-empty open set $U \subset \mathbb{C}^d$ such that for all $\boldsymbol{x} \in U$ the series

$$f(P, \boldsymbol{x}) = \sum_{m \in P \cap \mathbb{Z}^d} \boldsymbol{x}^m$$

converges absolutely and uniformly on compact subsets of $U$ to a rational function in $\boldsymbol{x}$. It turns out that this rational function can be efficiently computed as long as the dimension $d$ of the ambient space is fixed in advance.

The following result was proved, essentially, in [3] although the formal statement and better complexity bounds did not appear until [4].

**Theorem 2.2.** *Let us fix $d$. Then there exists a polynomial time algorithm, which, for a rational polyhedron $P \subset \mathbb{R}^d$ without lines defined by inequalities* (2.1) *computes the generating function*

$$f(P, \boldsymbol{x}) = \sum_{m \in P \cap \mathbb{Z}^d} \boldsymbol{x}^m$$

*in the form*

$$f(P, \boldsymbol{x}) = \sum_{i \in I} \varepsilon_i \frac{\boldsymbol{x}^{v_i}}{(1 - \boldsymbol{x}^{u_{i1}}) \ldots (1 - \boldsymbol{x}^{u_{id}})}, \tag{2.3}$$

*where $\varepsilon_i \in \{-1, 1\}$, $v_i, u_{ij} \in \mathbb{Z}^d$, and $u_{ij} \neq 0$ for all $i$, $j$.*

The complexity of the algorithm is $\mathcal{L}^{O(d)}$, where $\mathcal{L}$ is the input size of $P$ defined by (2.2). In particular, the number $|I|$ of terms in (2.3) is $\mathcal{L}^{O(d)}$, which is why we call (2.3) a *short rational function*.

Rational cones play the crucial role in the proof of Theorem 2.2.

**2.1. Rational cones.** A non-empty rational polyhedron $K$ is called a *rational cone* if for every $x \in K$ and $\lambda \geq 0$ we have $\lambda x \in K$. We are interested in *pointed* rational cones, that is, cones not containing lines (equivalently, cones for which 0 is the vertex). A basic example of a pointed rational cone is provided by the non-negative orthant $\mathbb{R}^d_+$ consisting of the points with non-negative coordinates. The generating function for the set of integer points in $\mathbb{R}^d_+$ is a multiple geometric series

$$f(\mathbb{R}^d_+, \boldsymbol{x}) = \sum_{m \in \mathbb{Z}^d_+} \boldsymbol{x}^m = \prod_{i=1}^d \frac{1}{1 - x_i}.$$

A *unimodular cone* $K$ is the set of non-negative linear combinations of a given basis $u_1, \ldots, u_d$ of the lattice $\mathbb{Z}^d$. Up to an integral change of coordinates, a unimodular cone $K$ looks like the non-negative orthant $\mathbb{R}^d_+$. Consequently, the generating function for the set of integer points in $K$ is a multiple geometric series

$$f(K, \boldsymbol{x}) = \sum_{m \in K \cap \mathbb{Z}^d} \boldsymbol{x}^m = \prod_{i=1}^d \frac{1}{1 - \boldsymbol{x}^{u_i}}.$$

It is well known that any rational cone $K$ can be subdivided into unimodular cones, cf., for example, Section 2.6 of [16]. However, even for $d = 2$, the number of the unimodular cones may have to be exponentially large in the input size: consider the cone $K \subset \mathbb{R}^2$ spanned by $(1, 0)$ and $(1, n)$ for a positive integer $n$. Nevertheless, where exists a computationally efficient procedure for constructing a more general *decomposition* of a rational cone into unimodular cones.

**Definition 2.3.** For a set $A \subset \mathbb{R}^d$, let $[A]\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ be the indicator of $A$ defined by

$$[A](x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Let $\mathcal{P}(\mathbb{Q}^d)$ be the vector space (over $\mathbb{C}$) spanned by the indicators $[P]$ of rational polyhedra $P \subset \mathbb{R}^d$. We call $\mathcal{P}(\mathbb{Q}^d)$ the *algebra of rational polyhedra*. The vector space $\mathcal{P}(\mathbb{Q}^d)$ possesses an interesting and useful algebra structure, cf. [26], which we do not discuss here.

The idea is to write the indicator $[K]$ of a given rational cone $K \subset \mathbb{R}^d$ as a linear combination of indicators of unimodular cones. For $d = 2$ such an efficient procedure has long been known via the *continued fractions* method, cf., for example, [22]. We give a simple example below.

Suppose that $K \subset \mathbb{R}^2$ is the cone spanned by vectors $(1, 0)$ and $(31, 164)$. Writing the continued fraction expansion, we obtain

$$\frac{164}{31} = 5 + \cfrac{1}{3 + \cfrac{1}{2 + \cfrac{1}{4}}} \quad ,$$

so we write $164/31 = [5; 3, 2, 4]$. Next, we compute the *convergents*

$$[5; 3, 2] = 5 + \cfrac{1}{3 + \cfrac{1}{2}} = \frac{37}{7}, \quad [5; 3] = 5 + \frac{1}{3} = \frac{16}{3}, \quad \text{and} \quad [5] = \frac{5}{1}$$

and notice that

$$[K] = [K_0] - [K_1] + [K_2] - [K_3] + [K_4],$$

where $K_0$ is spanned by $(1, 0)$ and $(0, 1)$, $K_1$ is spanned by $(0, 1)$ and $(1, 5)$, $K_2$ is spanned by $(1, 5)$ and $(3, 16)$, $K_3$ is spanned by $(3, 16)$ and $(7, 37)$, and $K_4$ is spanned by $(7, 37)$ and $(31, 164)$. Since $K_i$ turn out to be unimodular for $i = 0, 1, 2, 3, 4$, we get the short rational function expression

$$f(K, \boldsymbol{x}) = \frac{1}{(1 - x_1)(1 - x_2)} - \frac{1}{(1 - x_2)(1 - x_1 x_2^5)} + \frac{1}{(1 - x_1 x_2^5)(1 - x_1^3 x_2^{16})} - \frac{1}{(1 - x_1^3 x_2^{16})(1 - x_1^7 x_2^{37})} + \frac{1}{(1 - x_1^7 x_2^{37})(1 - x_1^{31} x_2^{164})}.$$

A polynomial time algorithm for computing a unimodular cone decomposition in any (fixed in advance) dimension $d$ was suggested in [3]. Using triangulations, it is not hard to reduce the case of an arbitrary rational cone to that of a *simple rational cone* $K \subset \mathbb{R}^d$

$$K = \left\{ \sum_{i=1}^{d} \lambda_i u_i : \ \lambda_i \geq 0 \right\}$$

spanned by linearly independent vectors $u_1, \ldots, u_d \in \mathbb{Z}^d$, which may not, however, constitute a basis of the lattice $\mathbb{Z}^d$. As a measure of how far is $K$ from being unimodular, we introduce the *index* $\mathrm{ind}(K)$ of $K$ as the index of the sublattice generated by $u_1, \ldots, u_d$ in the ambient lattice $\mathbb{Z}^d$. Thus $\mathrm{ind}(K)$ is a positive integer and $\mathrm{ind}(K) = 1$ if and only if $K$ is a unimodular cone.

Let us consider the parallelepiped

$$\Pi = \left\{ \sum_{i=1}^{d} \lambda_i u_i : \ |\lambda_i| \leq \mathrm{ind}^{-1/d}(K) \text{ for } i = 1, \ldots, d \right\}.$$

Then $\Pi$ is a convex body symmetric about the origin and $\mathrm{vol}\, \Pi = 2^d$. Therefore, by the Minkowski Theorem there is a non-zero point $w \in \Pi \cap \mathbb{Z}^d$, cf., for example, Section VII.3 of [5]. Moreover, such a point $w$ can be constructed in polynomial time as long as the dimension $d$ is fixed, cf. Section 6.7 of [17]. Replacing $w$ by $-w$ if needed, we can also ensure that $w$ lies in the same halfspace as $u_1, \ldots, u_d$. Let $K_i$ be the cone spanned by $u_1, \ldots, u_d$ with the vector $u_i$ replaced by $w$ and let $\varepsilon_i = 1$ or $\varepsilon_i = -1$ depending on whether this replacement preserves or reverses the orientation of the set $u_1, \ldots, u_d$ (we choose $\varepsilon_i = 0$ if we obtain a linearly dependent set). Then we observe that

$$\begin{aligned} &[K] = \sum_{i=1}^{d} \varepsilon_i [K_i] \ \pm \ \text{indicators of lower-dimensional cones, and} \\ &\mathrm{ind}(K_i) \leq \mathrm{ind}^{(d-1)/d}(K) \quad \text{if } \dim K_i = d. \end{aligned} \tag{2.4}$$

As we iterate the above procedure, on the $n$th step, we obtain a decomposition of the cone $K$ as a linear combination of at most $d^n$ cones $K_i$ (not counting smaller-dimensional cones) with

$$\mathrm{ind}(K_i) \leq (\mathrm{ind}(K))^{\left(\frac{d-1}{d}\right)^n}.$$

To ensure that all $K_i$ are unimodular, we can choose $n = O(d \log\log \mathrm{ind}(K))$, which results in a polynomial time algorithm for a fixed $d$.

To prove a weaker version of Theorem 2.2 (with $d$ replaced by $d+2$ in (2.3) and $\mathcal{L}^{O(d^2)}$ complexity) one can note that a rational polyhedron $P \subset \mathbb{R}^d$ without lines

can be represented as the section of a pointed rational cone $K \subset \mathbb{R}^{d+1}$ by the affine hyperplane $\xi_{d+1} = 1$. Consequently, we have

$$f(P, \boldsymbol{x}) = \frac{\partial}{\partial x_{d+1}} f(K, (\boldsymbol{x}, x_{d+1}))\big|_{x_{d+1}=0}. \tag{2.5}$$

**2.2. Using identities in the algebra of polyhedra.** The following remarkable result was proved by A. G. Khovanskii and A. V. Pukhlikov [23], and, independently, by J. Lawrence [25].

**Theorem 2.4.** *Let $\mathcal{P}(\mathbb{Q}^d)$ be the vector space spanned by the indicators of rational polyhedra and let $\mathbb{C}(\boldsymbol{x})$ be the vector space of rational functions in $d$ complex variables $\boldsymbol{x} = (x_1, \ldots, x_d)$. There exists a linear transformation $\mathcal{F}: \mathcal{P}(\mathbb{Q}^d) \longrightarrow \mathbb{C}(\boldsymbol{x})$ such that the following holds.*

(1) *If $P \subset \mathbb{R}^d$ is a rational polyhedron with a vertex then $\mathcal{F}([P]) = f(P, \boldsymbol{x})$, where $f(P, \boldsymbol{x})$ is the rational function defined as the sum of the series*

$$\sum_{m \in P \cap \mathbb{Z}^d} \boldsymbol{x}^m$$

*when the series converges absolutely.*

(2) *If $P \subset \mathbb{R}^d$ is a rational polyhedron without vertices then $\mathcal{F}([P]) = 0$.*

*Proof.* Let us fix a decomposition

$$\mathbb{R}^d = \sum_{i \in I} \alpha_i [Q_i] \tag{2.6}$$

for some rational polyhedra $Q_i$ with vertices and some numbers $\alpha_i$. Multiplying (2.6) by $[P]$, we get

$$[P] = \sum_{i \in I} \alpha_i [P \cap Q_i], \tag{2.7}$$

from which we deduce that $\mathcal{P}(\mathbb{Q}^d)$ is spanned by indicators of rational polyhedra with vertices.

Suppose that we have a linear relation

$$\sum_{j \in J} \beta_j [P_j] = 0 \tag{2.8}$$

for some polyhedra $P_j$ with vertices. Multiplying (2.8) by $[Q_i]$, we get

$$\sum_{j \in J} \beta_j [P_j \cap Q_i] = 0.$$

Since $Q_i$ has a vertex and $P_j \cap Q_i \subset Q_i$, there exists a non-empty open set $U_i \subset \mathbb{C}^d$ such that for all $\boldsymbol{x} \in U_i$ all the series defining $f(P_j \cap Q_i, \boldsymbol{x})$ converge absolutely and uniformly on compact subsets of $U_i$. Therefore, we must have

$$\sum_{j \in J} \beta_j f(P_j \cap Q_i, \boldsymbol{x}) = 0 \quad \text{for all } i \in I.$$

Similarly, from (2.7) we get

$$f(P_j, \boldsymbol{x}) = \sum_{i \in I} \alpha_i f(P_j \cap Q_i, \boldsymbol{x}) \quad \text{for all } j \in J.$$

Combining the last two equations, we conclude that

$$\sum_{j \in j} \beta_j f(P_j, \boldsymbol{x}) = \sum_{i \in I, j \in J} \alpha_i \beta_j f(P_j \cap Q_i, \boldsymbol{x}) = 0. \tag{2.9}$$

Thus a linear dependence (2.8) among indicators of rational polyhedra $P_j$ with vertices implies the corresponding linear dependence (2.9) among the generating functions $f(P_j, \boldsymbol{x})$. Therefore, the correspondence

$$[P] \longmapsto f(P, \boldsymbol{x})$$

extends to a linear transformation $\mathcal{F}: \mathcal{P}(\mathbb{Q}^d) \longrightarrow \mathbb{C}(\boldsymbol{x})$. It remains to show that $\mathcal{F}([P]) = 0$ if $P$ is a rational polyhedron with a line.

We observe that if $P' = P + u$ is a translation of $P$ by a lattice vector $u$, we must have $f(P', \boldsymbol{x}) = \boldsymbol{x}^u f(P, \boldsymbol{x})$ for all rational polyhedra $P$ with vertices. By linearity, we must have $\mathcal{F}([P + u]) = \boldsymbol{x}^u \mathcal{F}([P])$ for all rational polyhedra $P$. However, if $P$ contains a line then there is a vector $u \in \mathbb{Z}^d \setminus \{0\}$ such that $P + u = P$. Therefore, we must have $\mathcal{F}([P]) = 0$ for $P$ with a line. □

Theorem 2.4 provides a powerful tool for computing the generating function of the set of integer points in a rational polyhedron. The following "duality trick" going back to the seminal paper of M. Brion [11] turns out to be particularly useful.

Let $\langle \cdot, \cdot \rangle$ be the standard scalar product in $\mathbb{R}^d$ and let $K \subset \mathbb{R}^d$ be a cone. The cone

$$K^* = \{x \in \mathbb{R}^d : \langle x, y \rangle \geq 0 \text{ for all } y \in K\}$$

is called the *dual* to $K$. It is easy to see that if $K$ is rational (resp. unimodular) cone then $K^*$ is a rational (resp. unimodular) cone, and that if $K$ contains a line (resp. lies in a proper subspace of $\mathbb{R}^d$) then $K^*$ lies in a proper subspace of $\mathbb{R}^d$ (resp. contains a line). A standard duality argument implies that $(K^*)^* = K$ for closed convex cones $K$. A less obvious observation is that duality preserves linear relations among indicators of closed convex cones:

$$\sum_{i \in I} \alpha_i [K_i] = 0 \quad \text{implies} \quad \sum_{i \in I} \alpha_i [K_i^*] = 0,$$

see, for example, Section IV.1 of [5] for a proof.

Now, to compute the generating function $f(K, \boldsymbol{x})$ one can do the following. First, we compute the dual cone $K^*$, and, iterating (2.4), we compute unimodular cones $K_i$ and numbers $\varepsilon_i \in \{-1, 1\}$ such that

$$[K^*] \equiv \sum_{i \in I} \varepsilon_i [K_i] \quad \text{modulo indicators of lower-dimensional cones.}$$

Then, dualizing again, we get

$$[K] \equiv \sum_{i \in I} \varepsilon_i [K_i^*] \quad \text{modulo indicators of cones with lines.} \tag{2.10}$$

In view of Theorem 2.4, cones with lines can be ignored as far as generating functions are concerned. This gives us

$$f(K, \boldsymbol{x}) = \sum_{i \in I} \varepsilon_i f(K_i^*, \boldsymbol{x}).$$

Since $K_i^*$ are unimodular cones, this completes computation of $f(K, \boldsymbol{x})$. This trick allows us to reduce the complexity of the algorithm in Theorem 2.2 from $\mathcal{L}^{O(d^2)}$ to $\mathcal{L}^{O(d)}$, where $\mathcal{L}$ is the size of the input.

Another important identity is Brion's Theorem [11], which expresses the generating function of the set of integer points in $P$ as the sum of generating functions for the sets of integer points in the tangent (supporting) cones at the vertices of $P$. Namely, for a vertex $v$ of a polyhedron $P$ let us define the tangent cone $K_v$ as

$$K_v = \{x : \varepsilon x + (1 - \varepsilon)v \in P \text{ for all sufficiently small } \varepsilon > 0\}.$$

We note that $K_v$ is not a cone per se but rather a translation of the cone $K_v - v$.

**Theorem 2.5** (Brion's Theorem). *For a rational polyhedron $P$ we have*

$$f(P, \boldsymbol{x}) = \sum_{v} f(K_v, \boldsymbol{x}),$$

*where the sum is taken over all vertices of $P$ and the identity is understood as the identity among rational functions.*

Discovered by M. Brion [11], Theorem 2.5 started an avalanche of research. The original proof of Theorem 2.5 was based on algebro-geometric methods. Later, elementary proofs were discovered in [23] and [25]. One can deduce Theorem 2.5 from Theorem 2.4 and an elementary identity

$$[P] \equiv \sum_{v} [K_v] \quad \text{modulo indicators of polyhedra with lines,}$$

cf. Section VIII.4 of [5].

Theorem 2.5 together with the unimodular decomposition of Section 2.1 and the duality trick provide the proof of Theorem 2.2 as stated. Another advantage of using Theorem 2.5 is that it allows us to understand how the generating function $f(P, \boldsymbol{x})$ changes as the facets of $P$ move parallel to themselves so that the combinatorial structure of $P$ does not change. In this case, the tangent cones $K_v$ get translated by vectors linearly depending on the displacements of the facets of $P$. Writing $K_v$ as combinations of translated unimodular cones $K_i + v$ as in (2.10), we notice that as far as lattice points are concerned, a *rational* translation $K_i + v$ of a *unimodular* cone $K_i$ is equivalent to a certain *integer* translation $K_i + u$:

$$(K_i + v) \cap \mathbb{Z}^d = (K_i + u) \cap \mathbb{Z}^d \quad \text{for some } u \in \mathbb{Z}^d$$

and hence we have

$$f(K_i + v, \boldsymbol{x}) = f(K_i + u, \boldsymbol{x}) = \boldsymbol{x}^u f(K_i, \boldsymbol{x}).$$

If $K = \mathbb{R}^d_+$ then $u$ is obtained from $v$ by rounding up the coordinates to the nearest integer. The case of a general unimodular cone differs by a unimodular linear transformation, see [4] for details.

**2.3. Implementation.** The algorithm of Theorem 2.2 appears to be practical. First, it was implemented by J. De Loera et al. [12], who wrote the `LattE` (Lattice point enumeration) software package. The authors of `LattE` discovered that often the most practically efficient way to handle computations is to represent a polyhedron $P$ as a hyperplane section of a higher-dimensional cone as in (2.5) and then use the "dualized" decomposition (2.10). The package allows one to compute the number of integer points in a given rational polytope. Formally speaking, to compute the number $|P \cap \mathbb{Z}^d|$ of integer points in a given rational polytope $P$, we should substitute $\boldsymbol{x} = (1, \ldots, 1)$ into the rational function $f(P, \boldsymbol{x})$. However, we need to be careful since this particular value is a pole of every fraction in (2.3). Nevertheless, the substitution can be done efficiently, see Section 3.1 and [3], [4], [7], and [12] for details.

In addition, `LattE` allows one to compute the Ehrhart (quasi)-polynomial of a given rational polytope $P$, that is, to find a formula for the number of integer points in the dilated polytope $nP$, where $n$ is a positive integer, see also Section 6.1.

Testing whether a given rational polyhedron $P$ contains an integer point, or, equivalently, whether $f(P, \boldsymbol{x}) \not\equiv 0$ is a non-trivial problem related to the general *integer programming problem* of optimizing a given linear function on the set $P \cap \mathbb{Z}^d$. `LattE` package contains also an implementation of an integer programming algorithm based on rational functions $f(P, \boldsymbol{x})$.

Another implementation, called `barvinok`, was written by S. Verdoolaege, see [36]. Among other features, the implementation allows one to obtain closed explicit formulas for the number of integer points in a parametric polytope as a function of displacement parameters when the facets of the polytope move parallel to themselves, see Theorem 2.5 and the subsequent discussion.

There is an extensive literature devoted to the lattice point enumeration in polytopes, whether from algorithmic, structural, or application points of view. For the classical Ehrhart theory in the context of enumerative combinatorics, see [34] and [9] for a clever simplification of the proofs of the main results of the theory. For an approach featuring Dedekind sums and other analytic tools, see [8]. It does not seem to be possible to survey all the literature in the paper. In addition to already mentioned papers, we provide only a few references among many good papers which appeared after the survey [4].

Efficient counting in special situations with applications to computational questions in representation theory and network flows is discussed in [2]. For a recent advance connecting lattice point counting with algebraic geometry, see [29]. For a computationally efficient version of the Euler–Maclaurin formula, satisfying, in addition, some natural "local" conditions, see [10].

## 3. Operations on sets and generating functions

Motivated in part by Theorem 2.2, let us consider sets $S \subset \mathbb{Z}^d$ *defined* by their generating functions

$$f(S; \boldsymbol{x}) = \sum_{m \in S} \boldsymbol{x}^m$$

written as rational functions in the form

$$f(S; \boldsymbol{x}) = \sum_{i \in I} \varepsilon_i \frac{\boldsymbol{x}^{a_i}}{(1 - \boldsymbol{x}^{b_{i1}}) \dots (1 - \boldsymbol{x}^{b_{ik}})}. \tag{3.1}$$

Here $I$ is a finite set of indices, $\varepsilon_i \in \mathbb{Q}$, $a_i, b_{ij} \in \mathbb{Z}^d$, and $b_{ij} \neq 0$ for all $i, j$. To avoid ambiguity, we assume that either $S$ is finite, or, if $S$ is infinite, then there is a nonempty open set $U \subset \mathbb{C}^d$ such that the series defining $f(S; \boldsymbol{x})$ converges absolutely and uniformly on compact subsets of $U$ and for every fraction in (3.1) there is the Laurent series (multiple geometric series) expansion

$$\frac{\boldsymbol{x}^{a_i}}{(1 - \boldsymbol{x}^{b_{i1}}) \dots (1 - \boldsymbol{x}^{b_{ik}})} = \sum_{(\mu_1, \dots, \mu_k) \in \mathbb{Z}_+^k} \boldsymbol{x}^{a_i + \mu_1 b_{i1} + \dots + \mu_k b_{ik}}$$

in $U$.

To indicate the computational complexity level of our set $S$, we consider the two parameters *fixed* in formula (3.1): the number $d$ of variables and the number $k$ of binomials in the denominator of each fraction. Note that if we happen to have a smaller number of binomials in some fraction, we can formally "pad" it to $k$ by multiplying both the numerator and denominator of the fraction by some artificial binomials. Since $k$ is fixed, that would increase the length of the formula by a constant factor.

Next, we discuss what information about the set $S$ can be extracted from $f(S; \boldsymbol{x})$ given in the form of (3.1).

**3.1. Monomial substitutions and differentiation.** One piece of information we can get is the cardinality $|S|$ of a finite set $S$. To compute $|S|$, we would like to substitute $\boldsymbol{x} = (1, \ldots, 1)$ in (3.1), but this should be done carefully since this particular value of $\boldsymbol{x}$ is the pole of every single fraction in (3.1). The procedure is introduced in [3].

We choose a sufficiently generic vector $c \in \mathbb{Z}^d$, $c = (\gamma_1, \ldots, \gamma_d)$, so that $\langle c, b_{ij} \rangle \neq 0$ for all $i, j$. For a $\tau \in \mathbb{C}$, let

$$\boldsymbol{x}(\tau) = (e^{\tau\gamma_1}, \ldots, e^{\tau\gamma_d}).$$

Thus we want to compute

$$\lim_{\tau \longrightarrow 0} f(S; \boldsymbol{x}(\tau)).$$

Let us compute

$$\alpha_i = \langle c, a_i \rangle \quad \text{and} \quad \beta_{ij} = \langle c, b_{ij} \rangle.$$

Then

$$f(S; \boldsymbol{x}(\tau)) = \sum_{i \in I} \varepsilon_i \frac{e^{\alpha_i \tau}}{(1 - e^{\beta_{i1}\tau}) \ldots (1 - e^{\beta_{ik}\tau})}. \tag{3.2}$$

Next, we note that $f(S; \boldsymbol{x}(\tau))$ is a meromorphic function in $\tau$ and that we want to compute the constant term of its Laurent expansion in the neighborhood of $\tau = 0$. To do that, we deal with every fraction separately. We write each fraction of (3.2) as

$$\frac{e^{\alpha_i \tau}}{(1 - e^{\beta_{i1}\tau}) \ldots (1 - e^{\beta_{ik}\tau})} = \tau^{-k} e^{\alpha_i \tau} \prod_{j=1}^{k} g_{ij}(\tau), \quad \text{where } g_{ij}(\tau) = \frac{\tau}{1 - e^{\beta_{ij}\tau}}.$$

Now, each $g_{ij}(\tau)$ is an analytic function of $\tau$ and we compute is Taylor series expansion $p_{ij}(\tau)$ up to the $\tau^{k+1}$ term:

$$\frac{\tau}{1 - e^{\beta_{ij}\tau}} \equiv p_{ij}(\tau) \mod \tau^{k+1}.$$

Similarly, we compute a polynomial $q_i(\tau)$ such that

$$e^{\alpha_i \tau} \equiv q_i(\tau) \mod \tau^{k+1}.$$

Finally, successively multiplying polynomials $\mod \tau^{k+1}$ we compute the polynomial $h_i(\tau)$ with $\deg h_i \leq k$ such that

$$q_i \, p_{i1} \ldots p_{ik} \equiv h_i \mod \tau^{k+1}.$$

Letting

$$h(\tau) = \sum_{i \in I} h_i(\tau),$$

we conclude that the coefficient of $\tau^k$ in $h(\tau)$ is the desired value of (3.2) at $\tau = 0$ and hence is the value $f(S; \boldsymbol{x})$ at $\boldsymbol{x} = (1, \ldots, 1)$. We note that the procedure has a

polynomial time complexity even if both $k$ and $d$ are allowed to vary and if we allow different numbers $k_i \leq k$ of binomials in different fractions of (3.1).

A more general operation which can be computed in polynomial time is that of a *monomial substitution*. Let $f(\boldsymbol{x})$ be an expression of the type (3.1). Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be a new set of variables, let $l_1, \ldots, l_d \in \mathbb{Z}^n$ be vectors, and let $\phi\colon \mathbb{C}^n \longrightarrow \mathbb{C}^d$ be the transformation defined by

$$(z_1, \ldots, z_n) \longmapsto (x_1, \ldots, x_d) \quad \text{where } x_i = \boldsymbol{z}^{l_i}.$$

If the image $\phi(\mathbb{C}^n)$ does not lie in the set of poles of $f$, one can define a rational function $g(\boldsymbol{z}) = f(\phi(\boldsymbol{z}))$. Function $g$ can be computed in polynomial time in the form

$$g(\boldsymbol{z}) = \sum_{i \in I'} \delta_i \frac{\boldsymbol{z}^{q_i}}{(1 - \boldsymbol{z}^{b_{i1}}) \ldots (1 - \boldsymbol{z}^{b_{ik_i}})},$$

where $\delta_i \in \mathbb{Q}$, $q_i, b_{ij} \in \mathbb{Z}^n$, $b_{ij} \neq 0$ for all $i$, $j$ and $k_i \leq k$ for all $i \in I'$.

The case of $l_1 = \cdots = l_d = 0$ corresponds to the case of $\boldsymbol{x} = (1, \ldots, 1)$ considered above. As above, the general case of a monomial substitution is handled by a one-parametric perturbation and computation with univariate polynomials. Details can be found in [7] (the assumption that $k$ is fixed in advance is not needed there).

The operation of monomial substitution has the following geometric interpretation. Let $T\colon \mathbb{R}^d \longrightarrow \mathbb{R}^n$ be the linear transformation whose matrix in the standard bases consists of the integer column vectors $l_1, \ldots, l_d$. Let $S \subset \mathbb{Z}^d$ be a set and suppose that for all $m \in T(S)$ the set $T^{-1}(m) \cap S$ is finite. The monomial substitution $x_i = \boldsymbol{z}^{l_i}$ into the generating function $f(S; \boldsymbol{x})$ produces the weighted generating function $g(\boldsymbol{z})$ of the image $T(S) \subset \mathbb{Z}^n$, where each monomial $\boldsymbol{z}^m$ for $m \in T(S)$ is counted with multiplicity $|T^{-1}(m) \cap S|$.

Another useful operation is that of differentiation. Let $p$ be a $d$-variate polynomial. We can write

$$\sum_{m \in S} p(m) \boldsymbol{x}^m = p\left(x_1 \frac{\partial}{\partial x_1}, \ldots, x_d \frac{\partial}{\partial x_d}\right) f(S; \boldsymbol{x}).$$

As long as $k$ is fixed in advance, the result can be computed in polynomial time in the form

$$\sum_{i \in I'} \delta_i \frac{\boldsymbol{x}^{q_i}}{(1 - \boldsymbol{x}^{b_{i1}})^{\gamma_{i1}} \ldots (1 - \boldsymbol{x}^{b_{ik}})^{\gamma_{ik}}},$$

where $\delta_i \in \mathbb{Q}$, $a_i, b_{ij} \in \mathbb{Z}^d$, $b_{ij} \neq 0$, and $\gamma_{ij}$ are non-negative integers such that $\gamma_{i1} + \cdots + \gamma_{ik} \leq k + \deg p$ for all $i$, see [6].

This observation is used in [6], see also [10] and [13].

One corollary of Theorem 2.2 is that we can efficiently perform set-theoretic operations (intersection, union, difference) of finite sets defined by (3.1). The following result is proved in [7].

**Theorem 3.1.** *Let us fix positive integers $d$ and $k$. Then there exists a polynomial time algorithm, which, for any two finite sets $S_1, S_2 \subset \mathbb{Z}^d$ given by their rational generating functions*

$$f(S_1; \boldsymbol{x}) = \sum_{i \in I_1} \alpha_i \frac{\boldsymbol{x}^{p_i}}{(1 - \boldsymbol{x}^{a_{i1}}) \dots (1 - \boldsymbol{x}^{a_{ik}})} \tag{3.3}$$

*and*

$$f(S_2; \boldsymbol{x}) = \sum_{i \in I_2} \beta_i \frac{\boldsymbol{x}^{q_i}}{(1 - \boldsymbol{x}^{b_{i1}}) \dots (1 - \boldsymbol{x}^{b_{ik}})} \tag{3.4}$$

*computes the generating function $f(S; \boldsymbol{x})$ of their intersection $S = S_1 \cap S_2$ in the form*

$$f(S; \boldsymbol{x}) = \sum_{i \in I} \gamma_i \frac{\boldsymbol{x}^{u_i}}{(1 - \boldsymbol{x}^{v_{i1}}) \dots (1 - \boldsymbol{x}^{v_{is}})},$$

*where $s \leq 2k$.*

*Proof.* The idea of the proof is to *linearize* the operation of intersection of sets. Suppose we have two Laurent series

$$g_1(\boldsymbol{x}) = \sum_{m \in \mathbb{Z}^d} \rho_{1m} \boldsymbol{x}^m \quad \text{and} \quad g_2(\boldsymbol{x}) = \sum_{m \in \mathbb{Z}^d} \rho_{2m} \boldsymbol{x}^m.$$

Let us define their *Hadamard product* $g_1(\boldsymbol{x}) \star g_2(\boldsymbol{x})$ as

$$g(\boldsymbol{x}) = \sum_{m \in \mathbb{Z}^d} \rho_m \boldsymbol{x}^m \quad \text{where } \rho_m = \rho_{1m} \rho_{2m}.$$

Then, clearly,

$$f(S_1 \cap S_2; \boldsymbol{x}) = f(S_1; \boldsymbol{x}) \star f(S_2; \boldsymbol{x}).$$

Without loss of generality, we assume that there is a non-empty open set $U \subset \mathbb{C}^d$ such that for all $\boldsymbol{x} \in U$ and every fraction of (3.3) and (3.4) we have the multiple geometric series expansions:

$$\frac{\boldsymbol{x}^{p_i}}{(1 - \boldsymbol{x}^{a_{i1}}) \dots (1 - \boldsymbol{x}^{a_{ik}})} = \sum_{(\mu_1, \dots, \mu_k) \in \mathbb{Z}_+^k} \boldsymbol{x}^{p_i + \mu_1 a_{i1} + \dots + \mu_k a_{ik}} \tag{3.5}$$

and

$$\frac{\boldsymbol{x}^{q_i}}{(1 - \boldsymbol{x}^{b_{i1}}) \dots (1 - \boldsymbol{x}^{b_{ik}})} = \sum_{(\nu_1, \dots, \nu_k) \in \mathbb{Z}_+^k} \boldsymbol{x}^{q_i + \nu_1 b_{i1} + \dots + \nu_k b_{ik}}. \tag{3.6}$$

As usual, we assume that for all $\boldsymbol{x} \in U$ the convergence in (3.5) and (3.6) is absolute and uniform on all compact subsets of $U$. To ensure that such a set $U$ indeed exists, we choose a sufficiently generic linear function $\ell \colon \mathbb{R}^d \longrightarrow \mathbb{R}$ and make sure that

$\ell(a_{ij}), \ell(b_{ij}) > 0$ for all $i, j$ by reversing, if necessary, the direction of $a_{ij}$ and $b_{ij}$ via the identity

$$\frac{1}{1-\boldsymbol{x}^{a}} = -\frac{\boldsymbol{x}^{-a}}{1-\boldsymbol{x}^{-a}}.$$

Here we use that $S_1$ and $S_2$ are finite so that $f(S_1; \boldsymbol{x})$ and $f(S_2; \boldsymbol{x})$ are, in fact, Laurent polynomials.

Since the Hadamard product is a bilinear operation on series, in order to compute $f(S_1; \boldsymbol{x}) \star f(S_2; \boldsymbol{x})$ it suffices to compute the Hadamard product of every pair of series (3.5) and (3.6).

In the space $\mathbb{R}^{2k}$ of $2k$-tuples $(x, y)$, where $x = (\xi_1, \ldots, \xi_k)$ and $y = (\eta_1, \ldots, \eta_k)$, let us introduce the polyhedron

$$Q_i = \left\{ (x, y) : \begin{array}{l} \xi_1, \ldots, \xi_k;\ \eta_1, \ldots, \eta_k \geq 0 \\ p_i + \xi_1 a_{i1} + \cdots + \xi_k a_{ik} = q_i + \eta_1 b_{i1} + \cdots + \eta_k b_{ik} \end{array} \right\} \tag{3.7}$$

and let $\mathbb{Z}^{2k} \subset \mathbb{R}^{2k}$ be the standard integer lattice.

Since the Hadamard product is bilinear and for monomials we have

$$\boldsymbol{x}^{m_1} \star \boldsymbol{x}^{m_2} = \begin{cases} \boldsymbol{x}^{m} & \text{if } m_1 = m_2 = m \\ 0 & \text{if } m_1 \neq m_2, \end{cases}$$

the Hadamard product of the series (3.5) and (3.6) can be expressed as the sum

$$\sum_{(m,n) \in Q_i \cap \mathbb{Z}^{2k}} \boldsymbol{x}^{p_i + \mu_1 a_{i1} + \cdots + \mu_k a_{ik}}, \tag{3.8}$$

where $m = (\mu_1, \ldots, \mu_k)$ and $n = (\nu_1, \ldots, \nu_k)$. On the other hand, (3.8) is obtained from the generating function $f(Q_i, \boldsymbol{z})$ with $\boldsymbol{z} = (z_1, \ldots, z_{2k})$ by the monomial substitution

$$z_i = \boldsymbol{x}^{a_i} \ \text{ for } i = 1, \ldots, k \quad \text{and} \quad z_i = 1 \ \text{ for } i = k+1, \ldots, 2k \tag{3.9}$$

and multiplication by $\boldsymbol{x}^{p_i}$.

We use Theorem 2.2 to compute $f(Q_i, \boldsymbol{z})$. The monomial substitution (3.9) can also be computed in polynomial time, cf. Section 3.1. □

Therefore, one can compute the generating functions of the union and difference:

$$f(S_1 \cup S_2; \boldsymbol{x}) = f(S_1; \boldsymbol{x}) + f(S_2; \boldsymbol{x}) - f(S_1 \cap S_2; \boldsymbol{x})$$

and

$$f(S_1 \setminus S_2; \boldsymbol{x}) = f(S_1; \boldsymbol{x}) - f(S_1 \cap S_2; \boldsymbol{x}).$$

Theorem 3.1 allows us to work with generating functions (3.1) directly as with data structures bypassing any more explicit descriptions of sets $S$ in question. Of course, there is a price to pay: with every set-theoretic operation, the complexity level of the set, the number $k$ of binomials in the denominator of each fraction in (3.1), doubles. From the definition (3.7) of $Q_i$ we can notice that in a sufficiently general position we will have $\dim Q_i = 2k-d$, so we would be able to choose $s = 2k-d$ in Theorem 3.1. Theorem 3.1 admits an extension to infinite sets $S_1$ and $S_2$ provided there is a non-empty open set $U \subset \mathbb{C}^d$ such that the multiple geometric series expansions (3.5) and (3.6) hold for all fractions in (3.3) and (3.4). K. Woods [38] used the construction of the Hadamard product to show that in any fixed dimension there is a polynomial time algorithm to check if a given integer is a period of the Ehrhart quasi-polynomial of a given rational polytope.

## 4. Beyond polyhedra: projections

There are other interesting sets admitting short rational generating functions (3.1). We start with examples.

**4.1. Integer semigroups.** Let $S$ be the semigroup generated by positive coprime integers $a_1$ and $a_2$, that is, the set of all non-negative integer combinations of $a_1$ and $a_2$:

$$S = \{\mu_1 a_1 + \mu_2 a_2 : \ \mu_1, \mu_2 \in \mathbb{Z}_+\}.$$

It is not hard to show that

$$f(S; x) = \frac{1 - x^{a_1 a_2}}{(1 - x^{a_1})(1 - x^{a_2})}$$

(the series defining $f(S; x)$ converges for all $|x| < 1$).

Let $S$ be the semigroup generated by positive coprime integers $a_1$, $a_2$, and $a_3$,

$$S = \{\mu_1 a_1 + \mu_2 a_2 + \mu_3 a_3 : \ \mu_1, \mu_2, \mu_3 \in \mathbb{Z}_+\}.$$

Then there exist positive integers $p_1$, $p_2$, $p_3$, $p_4$, and $p_5$, not necessarily distinct, such that

$$f(S; x) = \frac{1 - x^{p_1} - x^{p_2} - x^{p_3} + x^{p_4} + x^{p_5}}{(1 - x^{a_1})(1 - x^{a_2})(1 - x^{a_3})}.$$

This interesting result was rediscovered a number of times. It was explicitly stated by M. Morales [27]; the proof was not published though. Independently, the proof was rediscovered by G. Denham [14]. Both proofs are algebraic and based on the interpretation of $f(S; x)$ as the Hilbert series of a graded ring $\mathbb{C}[t^{a_1}, t^{a_2}, t^{a_3}]$. In this special case (a Cohen–Macaulay ring of codimension 2), the Hilbert series can

be computed via the Hilbert–Burch Theorem, cf. also [18]. Meanwhile, a combinatorial proof of a somewhat weaker result (up to 12 monomials in the numerator) independently appeared in [35].

The pattern breaks down for semigroups with $d \geq 4$ generators, meaning that if we choose the denominator of $f(S; x)$ in the form $(1-x^{a_1}) \dots (1-x^{a_d})$, the number of monomials in the numerator does not remain constant for a particular value of $d$, and, moreover, grows exponentially with the input size of $a_1, \dots, a_d$. As shown in [35], for $d = 4$ the number of the monomials in the numerator can grow as fast as $\min^{1/2}\{a_1, a_2, a_3, a_4\}$, whereas the input size is only about $\log(a_1 a_2 a_3 a_4)$.

Nevertheless, the generating function $f(S; x)$ admits a short rational function representation for any number $d$ of generators fixed in advance. The following result was proved in [7].

**Theorem 4.1.** *Let us fix $d$. Then there exists a positive integer $s = s(d)$ and a polynomial time algorithm, which, given positive integers $a_1, \dots, a_d$, computes the generating function $f(S; x)$ of the semigroup*

$$S = \left\{ \sum_{i=1}^{d} \mu_i a_i \,:\, \mu_1, \dots, \mu_d \in \mathbb{Z}_+ \right\}$$

*generated by $a_1, \dots, a_d$ in the form*

$$f(S; x) = \sum_{i \in I} \alpha_i \frac{x^{p_i}}{(1 - x^{b_{i1}}) \dots (1 - x^{b_{is}})}, \tag{4.1}$$

*where $\alpha_i \in \mathbb{Q}$, $p_i, b_{ij} \in \mathbb{Z}$ and $b_{ij} \neq 0$ for all $i, j$.*

In particular, for any fixed $d$, the number $|I|$ of fractions in (4.1) is bounded by a polynomial in the input size, that is, in $\log(a_1 \dots a_d)$.

Theorem 4.1 is obtained as a corollary of a more general result that the *projection* of the set of integer points in a rational polytope of a fixed dimension admits a short rational generating function [7].

**Theorem 4.2.** *Let us fix $d$. Then there exists a number $s = s(d)$ and a polynomial time algorithm, which, given a rational polytope $P$ and a linear transformation $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that $T(\mathbb{Z}^d) \subset \mathbb{Z}^k$, computes the generating function $f(S; \boldsymbol{x})$ for $S = T(P \cap \mathbb{Z}^d)$, $S \subset \mathbb{Z}^k$, in the form*

$$f(S; \boldsymbol{x}) = \sum_{i \in I} \frac{\boldsymbol{x}^{p_i}}{(1 - \boldsymbol{x}^{b_{i1}}) \dots (1 - \boldsymbol{x}^{b_{is}})}, \tag{4.2}$$

*where $\alpha_i \in \mathbb{Q}$, $p_i, b_{ij} \in \mathbb{Z}^k$ and $b_{ij} \neq 0$ for all $i, j$.*

One can observe that Theorem 4.1 is a corollary of Theorem 4.2. Indeed, let $T : \mathbb{R}^d \longrightarrow \mathbb{R}$ be the linear transformation defined by

$$T(\xi_1, \dots, \xi_d) = a_1 \xi_1 + \dots + a_d \xi_d.$$

Then the semigroup $S$ generated by $a_1, \dots, a_d$ is the image $S = T(\mathbb{Z}_+^d)$ of the set $\mathbb{Z}_+^d$ of integer points in the rational polyhedron $\mathbb{R}_+^d \subset \mathbb{R}^d$. The polyhedron $\mathbb{R}_+^d$ is unbounded, so Theorem 4.2 cannot be applied immediately. However, it is not hard to show that $S \subset \mathbb{Z}_+$ stabilizes after a while (if $a_1, \dots, a_d$ are coprime then $S$ includes all sufficiently large positive integers). Thus only the initial interval of $S$ is of interest, to get which we replace $\mathbb{R}_+^d$ by a sufficiently large simplex

$$P = \left\{(\xi_1, \dots, \xi_d) : \sum_{i=1}^{d} \xi_i \leq t \text{ and } \xi_i \geq 0 \text{ for } i = 1, \dots, d\right\},$$

see [7] for details.

We sketch the proof of Theorem 4.2 below.

Without loss of generality we assume that $\dim \ker T = d - k$. The proof then proceeds by induction on $d - k$. If $d = k$ we are in the situation of Theorem 2.2. We note that for any $k$ and $d$, if the restriction $T : P \cap \mathbb{Z}^d \longrightarrow S$ is one-to-one, we can compute the generating function $f(S; \boldsymbol{x})$ from that of the set $P \cap \mathbb{Z}^d$ using an appropriate monomial substitution, cf. Section 3.1. Otherwise, the monomial substitution will account for each point $m \in S$ with the multiplicity equal to the number of the points in $P \cap \mathbb{Z}^d$ mapped onto $m$. Thus our goal is to eliminate multiplicities.

The case of $d = k + 1$ illuminates some of the ideas used in the proof for an arbitrary $d - k$. Suppose that

$$T : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}^k, \quad (\xi_1, \dots, \xi_{k+1}) \longmapsto (\xi_1, \dots, \xi_k)$$

is the projection (this is a sufficiently general case). Let $\hat{S} = P \cap \mathbb{Z}^{k+1}$ and let us consider the restriction $T : \hat{S} \longrightarrow S$. Then, for every point $m \in S$, the preimage $T^{-1}(m) \subset \hat{S}$ is the set of integer points in the interval $T^{-1}(m) \cap P$ which all agree in their first $k$ coordinates and disagree in the last coordinate. Let $e_{k+1}$ be the last basis vector and let us consider

$$Y = \hat{S} \setminus (\hat{S} + e_{k+1}).$$

In words: we subtract from $\hat{S}$ its translation by 1 in the last coordinate.

Then the restriction $T : Y \longrightarrow S$ is one-to-one since the preimage $T^{-1}(m) \subset Y$ consists of the single point in $T^{-1}(m) \subset \hat{S}$ with the smallest last coordinate. Now, $\hat{S}$ is the set of integer points in a rational polytope and we compute its generating function using Theorem 2.2. Then we compute the generating function of $Y$ using Theorem 3.1. Finally, we obtain $f(S; \boldsymbol{x})$ by substituting $x_{k+1} = 1$ in the generating function $f(Y; (\boldsymbol{x}, x_{k+1}))$, cf. Section 3.1.

Let us consider the case of general $k$ and $d$. Let $\mathrm{pr} : \mathbb{Z}^{k+1} \longrightarrow \mathbb{Z}^k$ be the natural projection, $\mathrm{pr}(\mu_1, \dots, \mu_{k+1}) = (\mu_1, \dots, \mu_k)$. Let $\hat{T} : \mathbb{Z}^d \longrightarrow \mathbb{Z}^{k+1}$ be a linear transformation which is a lifting of $T$ so that $\mathrm{pr}(\hat{T}(m)) = T(m)$ for all $m \in \mathbb{Z}^d$. We define $\hat{S} = \hat{T}(S)$, $\hat{S} \subset \mathbb{Z}^{k+1}$, and consider the restriction

$$\mathrm{pr} : \hat{S} \longrightarrow S.$$

For every $m \in S$ the preimage $\mathrm{pr}^{-1}(m) \subset \hat{S}$ consists of the points which differ in their last coordinate only. Suppose that we managed to construct $\hat{T}$ in such a way that the set $\mathrm{pr}^{-1}(m) \subset \hat{S}$ has *small gaps*, meaning that there exists a constant $l = l(d)$ such that if there are two points in $\mathrm{pr}^{-1}(m)$ whose $(k+1)$st coordinates differ by more than $l$, there must be a point in $\mathrm{pr}^{-1}(m)$ lying strictly between them.

In this case, we compute $f(S; \boldsymbol{x})$ as follows. Let us define

$$Y = \hat{S} \setminus \bigcup_{j=1}^{l} (\hat{S} + j e_{k+1}).$$

In words: we subtract from $\hat{S}$ its $l$ translates by $1, \ldots, l$ in the last coordinate. Because of the small gap property, the restriction $\mathrm{pr}: Y \longrightarrow S$ is one-to-one: now, the preimage $\mathrm{pr}^{-1}(m) \subset Y$ consists of the single point in $\mathrm{pr}^{-1}(m) \subset \hat{S}$ with the smallest last coordinate. Using the induction hypothesis, we compute the generating function of $\hat{S}$. Then, applying Theorem 3.1 $l$ times, we compute the generating function of $Y$. Finally, $f(S; \boldsymbol{x})$ is obtained from $f(Y; (\boldsymbol{x}, x_{k+1}))$ by the substitution $x_{k+1} = 1$, see Section 3.1.

In general, we cannot construct a lifting $\hat{T}$ with the small gap property but the next best thing is possible. Namely, we can construct in polynomial time a decomposition $\mathbb{R}^k = \bigcup_i Q_i$ of $\mathbb{R}^k$ into a union of non-overlapping rational polyhedra $Q_i$ such that for each piece $S_i = S \cap Q_i$ a lifting $\hat{T}_i$ with the small gap property indeed exists. The generating functions $f(S_i; \boldsymbol{x})$ are computed as above and then patched together into a single generating function $f(S; \boldsymbol{x})$. The construction of such polyhedra $Q_i$ and liftings $\hat{T}_i$ is based on the results of [21] and [20]. The main tool is the following *Flatness Theorem*, see, for example, Section 6.7 of [17] or Section VII.8 of [5].

**Theorem 4.3** (Flatness Theorem). *For each dimension $d$ there exists a constant $\omega(d)$ with the following property: if $V$ is a $d$-dimensional real vector space, $\Lambda \subset V$ is a lattice of rank $d$, $\Lambda^* \subset V^*$ is the reciprocal lattice, and $K \subset V$ is a convex compact set with non-empty interior such that $K \cap \Lambda = \emptyset$ then there is an $\ell \in \Lambda^* \setminus \{0\}$ such that*

$$\max_{x \in K} \ell(x) - \min_{x \in K} \ell(x) \leq \omega(d). \tag{4.3}$$

In words: a lattice-free convex body is flat in some lattice direction. The number in the left hand side of (4.3) is called the *width of $K$ with respect to $\ell$* and denoted $\mathrm{width}(K, \ell)$. The infimum of $\mathrm{width}(K, \ell)$ over all $\ell \in \Lambda^*$ is called the *lattice width of $K$* and denoted $\mathrm{width}(K)$. A simple and crucial observation relating the lattice width and the small gap property is that if for $\ell \in \Lambda^*$ we have $\mathrm{width}(K, \ell) \leq \gamma\, \mathrm{width}(K)$ then the gaps between the consecutive integers in the set $\ell(K \cap \Lambda)$ do not exceed $\gamma\omega(d)$.

We go back to finish the sketch of the proof of Theorem 4.2. Let $\Lambda = \mathbb{Z}^k \cap \ker(T)$ be the lattice in $\ker(T)$. For $y \in \mathbb{R}^d$, let $P_y = P \cap T^{-1}(x)$ be the fiber of the polytope $P$ over $x$. We will measure the lattice width of $P_y$ with respect to $\Lambda$. The results of [21]

and [20] allow us to construct a polyhedral decomposition $\mathbb{R}^k = \bigcup_i Q_i$ and vectors $\ell_i \in \Lambda^*$ such that for all $y \in Q_i$ we have either $\text{width}(P_y, \ell_i) \leq 2\,\text{width}(P_y)$ or $\text{width}(P_y, \ell_i) \leq 1$. We then define

$$\hat{T}_i(x) = (T(x), \ell_i(x)) \quad \text{if } T(x) \in Q_i.$$

This completes the sketch of proof of Theorem 4.2.

**4.2. Applications.** Theorem 4.1 implies polynomial time solvability of a variety of problems about integer semigroups. Suppose that the generators $a_1, \ldots, a_d$ are coprime. As is known, all sufficiently large integers lie in the semigroup $S$ generated by $a_1, \ldots, a_d$. In the situation when the number $d$ of generators is fixed, R. Kannan [20] constructed a polynomial time algorithm to compute the largest integer not in $S$. Theorem 4.1 implies that one can compute in polynomial time the number of positive integers not in $S$, the number of integers in $S$ belonging to a particular interval, etc.

Unlike the algorithm of Theorem 2.2, the algorithms of Theorems 4.1 and 4.2 seem to be unimplementable at the moment. Indeed, the way Theorem 4.2 is proved gives $s = d^{\Omega(d)}$ at best and, similarly, in Theorem 4.1. It is not clear at the moment whether a smaller value of $s$ is possible.

In Theorem 4.1, apart from $d = 1, 2, 3$, the value of $d = 4$ seems to indicate a possibility of a "special treatment". The approach of [33] combined with the continued fraction method, see Section 2.1, may lead to a practically efficient algorithm to compute $f(S; x)$.

Theorem 4.2 implies that some other interesting sets admit short rational generating functions. One class of such sets consists of the Hilbert bases of rational cones. Let $K \subset \mathbb{R}^d$ be a pointed rational cone. The set $S \subset K \cap \mathbb{Z}^d$, $0 \notin S$, is called the (minimal) *Hilbert basis* of the semigroup $K \cap \mathbb{Z}^d$ if every point in $K \cap \mathbb{Z}^d$ can be represented as a sum of some points in $S$ and if no point in $S$ is a sum of other points in $S$. In other words, $S$ consists of the points in $K \cap \mathbb{Z}^d$ that cannot be written as a sum of non-zero points in $K \cap \mathbb{Z}^d$. Theorem 4.2 implies that as long as the dimension $d$ remains fixed, given a rational cone $K$, the generating function $f(S; \boldsymbol{x})$ can be computed in polynomial time as a short rational function of the type (3.1). Consequently, the number $|S|$ of points in the Hilbert basis of $K \cap \mathbb{Z}^d$ can be computed in polynomial time.

To deduce this result from Theorem 4.2, let $Q \subset K$ be a rational polyhedron containing all integer points in $K$ except 0 (to get $Q$ from $K$, we cut the vertex of $K$ by a hyperplane), let $P = Q \times Q \subset \mathbb{R}^d \oplus \mathbb{R}^d = \mathbb{R}^{2d}$ and let $T$ be the projection $P \longmapsto K$, $T(x, y) = x + y$. Then the Hilbert basis $S$ is the complement in $Q \cap \mathbb{Z}^d$ of the image $T(P \cap \mathbb{Z}^{2d})$. The obstacle that the polyhedron $Q$ is not bounded, so Theorem 4.2 cannot be applied immediately, can be easily fixed since only the "initial part" of the semigroup $K \cap \mathbb{Z}^d$ is of interest, see [7].

Another class of sets allowing short rational generating functions via Theorem 4.2 are the *test sets* in integer programming, see [30].

It should be noted that the short rational function description provides only very general characterization of the set. For example, many of the fine properties of test sets [30] do not seem to be picked up by rational generating functions and some empirically observed phenomena are still waiting for their explanation. For structural results (without complexity estimates) regarding $f(S; \boldsymbol{x})$, where $S$ is the projection of the set of integer points in a rational polyhedron, see [24].

## 5. Beyond projections: Presburger arithmetic

Let us consider formulas we can construct by using integer variables, operations of addition, subtraction, and multiplication by an integer constant (but not multiplication of two integer variables), comparison ($<$, $>$, $=$), Boolean operations ("and", "or", "not"), and quantifiers ($\forall$, $\exists$). The realm of such formulas is *Presburger arithmetic*. Thus the set $P \cap \mathbb{Z}^d$ of integer points in a rational polyhedron can be described by a quantifier-free formula of Presburger arithmetic: the set $P \cap \mathbb{Z}^d$ consists of the $d$-tuples of integer variables that satisfy a number of linear constraints with constant integer coefficients. Similarly, the projection $T(P \cap \mathbb{Z}^d)$ of the set of integer points in a polyhedron is described by a formula of Presburger arithmetic with existential quantifiers only (no quantifier alternations).

With a little work, Theorem 2.2 can be extended as follows. Let us fix the number $d$ of variables. Then there exists a polynomial time algorithm, which, given a quantifier-free formula $F$ of Presburger arithmetic, computes the generating function $f(S; \boldsymbol{x})$ of the set $S \subset \mathbb{Z}^d$ defined by $F$ as a rational function (2.3). Some routine precautions regarding convergence of the series defining $f(S; \boldsymbol{x})$, if $S$ is infinite, should be taken. The general case of a set defined by a quantifier-free formula $F$ reduces to that of the set integer points in a rational polyhedron by some more or less straightforward "cutting and pasting" of polyhedra. Since the dimension $d$ of the ambient space is fixed, this cutting and pasting can be performed in polynomial time.

Theorem 4.2 can be extended as follows. Let us fix the number of variables *and* the number of Boolean operations used. Then there exists a polynomial time algorithm, which, given a formula $F$ of Presburger arithmetic without quantifier alternations, computes the generating function $f(S; \boldsymbol{x})$ of the finite set $S \subset \mathbb{Z}^k$ defined by $F$ as a rational function (4.2). Note that here we have to fix not only the number of variables but also the number of Boolean operations. For example, unless $\mathbf{P} = \mathbf{NP}$ one cannot hope to compute the generating function of the projection of the set of integer points in a union of rational polytopes if the number of polytopes is allowed to vary, cf. Section 5.3 of [37] and [31].

One can ask whether the results can be extended even further. Let us fix the number of variables and the number of Boolean operations, making numerical constants essentially the only parameters of the formula. Is there a polynomial time algorithm which computes the generating function (3.1) of the (finite) set $S$ of points described by such a formula? This indeed seems very plausible, see the discussion in Chapter V

of [37]. Intuitively, such sets should have some "hidden periodicity" and short rational generating functions should reveal that periodicity. Besides, it seems hard to prove that a particular finite, but large, set $S \subset \mathbb{Z}^d$ does not admit a short rational generating function: if a particular candidate expression for $f(S; \boldsymbol{x})$ is not short, one can argue that we have not searched hard enough and that there is another, better candidate.

We mention that the result of R. Kannan [19] establishes polynomial time solvability of decision problems for formulas with not more than one quantifier alternation. If the number of variables is not fixed, the complexity of decision problems in Presburger arithmetic is double exponential by the result of M. Fischer and M. Rabin [15].

## 6. Concluding remarks

One can ask whether some of the technique discussed in this paper can be extended to lattice points satisfying some non-linear constraints. The answer seems to be "no". For example, lattice points in the standard Euclidean ball exhibit phenomena explained not by rational but rather by theta functions. Let

$$B_n = \{(\xi_1, \xi_2, \xi_3, \xi_4) : \ \xi_1^2 + \xi_2^2 + \xi_3^2 + \xi_4^2 \leq n\}$$

be the Euclidean ball of radius equal to $\sqrt{n}$. Jacobi's formula asserts that the number $|B_n \cap \mathbb{Z}^4| - |B_{n-1} \cap \mathbb{Z}^4|$ of integer points on the sphere of radius $\sqrt{n}$ is equal to

$$8 \sum_{4 \nmid r \mid n} r$$

(in words: eight times the sum of divisors of $n$ that are not divisible by four). One can then show ([1]) that if one can count points in a 4-dimensional ball efficiently (in polynomial time), one can factor integers efficiently (in randomized polynomial time).

We note also that lattice points in *irrational* polyhedra exhibit a very interesting behavior, see [32].

**6.1. Large dimensions.** Almost everywhere in this paper we assumed that the dimension $d$ of the ambient space is fixed in advance. But what if the dimension is allowed to grow? Given a rational polyhedron $P \subset \mathbb{R}^d$, it is an NP-hard problem to determine whether $P \cap \mathbb{Z}^d = \emptyset$ (even when $P$ is a rational simplex). Thus there is little hope to compute the generating function $f(P, \boldsymbol{x})$ in polynomial time. However, it appears that some interesting "residues" or "shadows" of $f(P, \boldsymbol{x})$ can be efficiently computed even when the dimension $d$ is allowed to grow, cf. [10] and [6].

The number $e(P) = |P \cap \mathbb{Z}^d|$ of integer points in a rational polyhedron is an example of a lattice invariant *valuation*, see [26]. That is, the map $P \longmapsto e(P)$ extends to a linear functional on the space spanned by the indicators $[P]$ of rational polyhedra, cf. Definition 2.3, and the linear functional is invariant under lattice shifts:

$e(P) = e(P+u)$, $u \in \mathbb{Z}^d$. One can ask if there is another lattice invariant valuation $\nu$ on rational polytopes which is efficiently computable in interesting cases and which, in some sense, approximates the counting valuation $e(P)$. For example, the volume vol $P$ may serve as the "0th" approximation to $e(P)$.

With every lattice invariant valuation $\nu$ one can associate the expression

$$\nu(nP) = \sum_{i=0}^{d} \nu_i(P; n) n^i, \tag{6.1}$$

where $nP$ is a dilation of $P$ by an integer factor $n$ and the coefficients $\nu_i(P; n)$ are quasi-periodic: $\nu_i(P; n + t) = \nu_i(P; n)$ provided $tP$ is a polytope with integer vertices, cf. [26]. In the case of the counting valuation $e$, the expression (6.1) is called the *Ehrhart quasi-polynomial* of $P$ and $e_d(P; n) = \text{vol}\, P$. As the $k$th approximation to the counting valuation $e$ we consider a lattice invariant valuation $\nu$ which agrees with $e$ in the $k+1$ highest terms:

$$\nu_i(P; n) = e_i(P; n) \quad \text{for} \quad i = d, d-1, \ldots, d-k.$$

A natural goal is to construct such a valuation $\nu$, which is computable in polynomial time (at least, in some interesting cases) for any $k$ fixed in advance.

Abstractly speaking, to define the counting valuation $e$, we have to choose a finite-dimensional real vector space $V$ and a lattice $\Lambda \subset V$. Then we define $e(P) = |P \cap \Lambda|$ for every polytope $P \subset V$ such that the vertices of $tP$ belong to $\Lambda$ for some integer $t$. Apparently, to make a canonical choice of $\nu$, we have to fix some additional structure in $V$. In [6] a canonical valuation $\nu$ is constructed for rational polytopes whose facets are parallel to hyperplanes from a given finite collection of hyperplanes. Valuation $\nu$ agrees with $e$ in the $k+1$ highest terms and for any fixed $k$ valuation $\nu$ is polynomially computable on polytopes with the number facets exceeding the dimension $d$ by not more than a constant fixed in advance (in particular, on rational simplices). In [10] a different canonical valuation $\mu$ is constructed provided a scalar product on $V$ is chosen. Valuation $\mu$ also agrees with $e$ on the $k+1$ highest terms and polynomially computable on the same class of polytopes.

## References


[1] Bach, E., Miller, G., Shallit, J., Sums of divisors, perfect numbers and factoring. *SIAM J. Comput.* **15** (1986), 1143–1154.

[2] Baldoni-Silva, W., De Loera, J. A., Vergne, M. Counting integer flows in networks. *Found. Comput. Math.* **4** (2004), 277–314.

[3] Barvinok, A.I., A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math. Oper. Res.* **19** (1994), 769–779.

[4] Barvinok, A., Pommersheim, J. E., An algorithmic theory of lattice points in polyhedra. In *New Perspectives in Algebraic Combinatorics* (Berkeley, CA, 1996–97). Math. Sci. Res. Inst. Publ. 38, Cambridge University Press, Cambridge 1999, 91–147.

[5] Barvinok, A., *A Course in Convexity*. Graduate Studies in Mathematics 54, American Mathematical Society, Providence, RI, 2002.

[6] Barvinok, A., Computing the Ehrhart quasi-polynomial of a rational simplex. *Math. Comput.* **75** (2006), 1449–1466.

[7] Barvinok, A., Woods, K., Short rational generating functions for lattice point problems. *J. Amer. Math. Soc.* **16** (2003), 957–979.

[8] Beck, M., Robins, S., *Computing the Continuous Discretely. Integer-point Enumeration in Polyhedra*. Undergraduate Texts in Mathematics, Springer-Verlag, Berlin, to appear.

[9] Beck, M., Sottile, F., Irrational proofs for three theorems of Stanley. Preprint, arXiv math.CO/0501359, 2005.

[10] Berline, N., Vergne, M., Local Euler-Maclaurin formula for polytopes. Preprint, arXiv math.CO/0507256, 2005.

[11] Brion, M. Points entiers dans les polyèdres convexes. *Ann. Sci. École Norm. Sup.* (4) **21** (1988), 653–663.

[12] De Loera, J. A., Hemmecke, R., Tauzer, J., Yoshida, R., Effective lattice point counting in rational convex polytopes. *J. Symbolic Comput.* **38** (2004), 1273–1302; see also http://www.math.ucdavis.edu/~latte/

[13] De Loera, J. A., Hemmecke, R., Köppe, M., Weismantel, R., Integer polynomial optimization in fixed dimension. *Math. Oper. Res.*, to appear.

[14] Denham, G., Short generating functions for some semigroup algebras. *Electron. J. Combin.* **10** (2003), Research Paper 36, 7 pp. (electronic).

[15] Fischer, M. J., Rabin, M. O., Super-exponential complexity of Presburger arithmetic. In *Complexity of Computation* (Proc. SIAM-AMS Sympos., New York, 1973), SIAM-AMS Proc. VII, Amer. Math. Soc., Providence, R.I., 1974, 27–41.

[16] Fulton, W., *Introduction to Toric Varieties*. Annals of Mathematics Studies 131, Princeton University Press, Princeton 1993.

[17] Grötschel, M., Lovász, L., Schrijver, A., *Geometric Algorithms and Combinatorial Optimization*. Second edition, Algorithms and Combinatorics 2, Springer-Verlag, Berlin 1993.

[18] Herzog, J., Generators and relations of abelian semigroups and semigroup rings. *Manuscripta Math.* **3** (1970), 175–193.

[19] Kannan, R., Test sets for integer programs, $\forall\exists$ sentences. In *Polyhedral Combinatorics* (Morristown, NJ, 1989), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 1, Amer. Math. Soc., Providence, RI, 1990, 39–47.

[20] Kannan, R., Lattice translates of a polytope and the Frobenius problem. *Combinatorica* **12** (1992), 161–177.

[21] Kannan, R., Lovász, L., Scarf, H. E., The shapes of polyhedra. *Math. Oper. Res.* **15** (1990), 364–380.

[22] Khinchin, A. Ya., *Continued Fractions*. The University of Chicago Press, Chicago, IL, London 1964.

[23] Khovanskii, A. G., Pukhlikov, A. V., The Riemann-Roch theorem for integrals and sums of quasipolynomials on virtual polytopes. *Algebra i Analiz* **4** (4) (1992), 188–216; English translation in *St. Petersburg Math. J.* **4** (4) (1993), 789–812.

[24] Khovanskii, A. G., Sums of finite sets, orbits of commutative semigroups and Hilbert functions. *Funktsional. Anal. i Prilozhen.* **29** (2) (1995), 36–50, 95; English translation in *Funct. Anal. Appl.* **29** (2) (1995), 102–112.

[25] Lawrence, J., Rational-function-valued valuations on polyhedra. In *Discrete and Computational Geometry* (New Brunswick, NJ, 1989/1990), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 6, Amer. Math. Soc., Providence, RI, 1991, 199–208.

[26] McMullen, P., Valuations and dissections. In *Handbook of Convex Geometry*, Vol. B, North-Holland, Amsterdam 1993, 933–988.

[27] M. Morales, Syzygies of monomial curves and a linear diophantine problem of Frobenius. Preprint, Max-Planck-Institut für Mathematik, Bonn 1986.

[28] Papadimitriou, C. H., *Computational Complexity*. Addison-Wesley Publishing Company, Reading, MA, 1994.

[29] Pommersheim, J., Thomas, H., Cycles representing the Todd class of a toric variety. *J. Amer. Math. Soc.* **17** (2004), 983–994.

[30] Scarf, H. E., Test sets for integer programs. In *Lectures on Mathematical Programming* (ISMP97, Lausanne, 1997), Math. Programming **79** (1997), no. 1-3, Ser. B, 355–368.

[31] Schöning, U., Complexity of Presburger arithmetic with fixed quantifier dimension. *Theory Comput. Syst.* **30** (1997), 423–428.

[32] Skriganov, M. M., Ergodic theory on SL($n$), Diophantine approximations and anomalies in the lattice point problem. *Invent. Math.* **132** (1998), 1–72.

[33] Shallcross, D., Neighbors of the origin for four by three matrices. *Math. Oper. Res.* **17** (1992), 608–614.

[34] Stanley, R. P., *Enumerative Combinatorics*. Vol. 1, corrected reprint of the 1986 original, Cambridge Studies in Advanced Mathematics 49, Cambridge University Press, Cambridge 1997.

[35] Székely L. A., Wormald, N. C., Generating functions for the Frobenius problem with 2 and 3 generators. *Math. Chronicle* **15** (1986), 49–57.

[36] Verdoolaege, S., Woods, K., Bruynooghe M., Cools R., Computation and manipulation of enumerators of integer projections of parametric polytopes. Preprint Katholieke Universiteit Leuven, Dept. of Computer Science, Report CW 392, 2005; see also http://www.kotnet.org/~skimo/barvinok/

[37] Woods, K. M., Rational generating functions and lattice point sets. Diss. University of Michigan, 2004.

[38] Woods, K., Computing the period of an Ehrhart quasipolynomial. *Electron J. Combin.* **12** (2005), Research paper 34, 12 pp. (electronic).

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1043, U.S.A.
E-mail: barvinok@umich.edu

# Rational and algebraic series in combinatorial enumeration

Mireille Bousquet-Mélou

**Abstract.** Let $\mathcal{A}$ be a class of objects, equipped with an integer size such that for all $n$ the number $a_n$ of objects of size $n$ is finite. We are interested in the case where the generating function $\sum_n a_n t^n$ is rational, or more generally algebraic. This property has a practical interest, since one can usually say a lot on the numbers $a_n$, but also a combinatorial one: the rational or algebraic nature of the generating function suggests that the objects have a (possibly hidden) structure, similar to the *linear structure* of words in the rational case, and to the *branching structure* of trees in the algebraic case. We describe and illustrate this combinatorial intuition, and discuss its validity. While it seems to be satisfactory in the rational case, it is probably incomplete in the algebraic one. We conclude with open questions.

**Mathematics Subject Classification (2000).** Primary 05A15; Secondary 68Q45.

**Keywords.** Enumerative combinatorics, generating functions, rational and algebraic power series, formal languages.

## 1. Introduction

The general topic of this paper is the enumeration of discrete objects (words, trees, graphs,…) and more specifically the *rational* or *algebraic* nature of the associated generating functions. Let $\mathcal{A}$ be a class of discrete objects equipped with a size:

$$\begin{array}{rccc} \text{size}: & \mathcal{A} & \to & \mathbb{N} \\ & A & \mapsto & |A|. \end{array}$$

Assume that for all $n$, the number $a_n$ of objects of size $n$ is finite. The *generating function of the objects of* $\mathcal{A}$, *counted by their size*, is the following formal power series in the indeterminate $t$:

$$A(t) := \sum_{n\geq 0} a_n t^n = \sum_{A\in\mathcal{A}} t^{|A|}. \tag{1}$$

To take a very simple example, if $\mathcal{A}$ is the set of words on the alphabet $\{a, b\}$ and the size of a word is its number of letters, then the generating function is $\sum_{n\geq 0} 2^n t^n = 1/(1-2t)$.

Generating functions provide both a tool for solving counting problems, and a concise way to encode their solution. Ideally, one would probably dream of finding a closed formula for the numbers $a_n$. But the world of mathematical objects would be extremely poor if this was always possible. In practise, one is usually happy with an

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

expression of the generating function $A(t)$, or even with a recurrence relation defining the sequence $a_n$, or a functional equation defining $A(t)$.

Enumerative problems arise spontaneously in various fields of mathematics, computer science, and physics. Among the most generous suppliers of such problems, let us cite discrete probability theory, the analysis of the complexity of algorithms [56], [44], and the discrete models of statistical physics, like the famous Ising model [5]. More generally, counting the objects that occur in one's work seems to answer a natural curiosity. It helps to understand the objects, for instance to appreciate how restrictive are the conditions that define them. It also forces us to get some understanding of the *structure* of the objects: an enumerative result never comes for free, but only after one has elucidated, at least partly, what the objects really are.

We focus in this survey on objects having a rational, or, more generally, algebraic generating function. Rational and algebraic formal power series are well-behaved objects with many interesting properties. This is one of the reasons why several classical textbooks on enumeration devote one or several chapters to these series [43], [74], [75]. These chapters give typical examples of objects with a rational [resp. algebraic] generating function (GF). After a while, the collection of these examples builds up a general picture: one starts thinking that yes, all these objects have something in common in their structure. At the same time arises the following question: do all objects with a rational [algebraic] GF look like that? In other words, what does it mean, what does it suggest about the objects when they are counted by a rational [algebraic] GF?

This question is at the heart of this survey. For each of the two classes of series under consideration, we first present a general family of enumerative problems whose solution falls invariably in this class. These problems are simple to describe: the first one deals with walks in a directed graph, the other with plane trees. Interestingly, these families of objects admit alternative descriptions in language theoretic terms: they correspond to *regular languages*, and to *unambiguous context-free languages*, respectively. The words of these languages have a clear recursive structure, which explains directly the rationality [algebraicity] of their GF.

The series counting words of a regular [unambiguous context-free] language are called $\mathbb{N}$-*rational* [$\mathbb{N}$-*algebraic*]. It is worth noting that a rational [algebraic] series with non-negative coefficients is not necessarily $\mathbb{N}$-rational [$\mathbb{N}$-algebraic]. Since we want to appreciate whether our two generic classes of objects are good representatives of objects with a rational [algebraic] GF, the first question to address is the following: do we always fall in the class of $\mathbb{N}$-rational [$\mathbb{N}$-algebraic] series when we count objects with a rational [algebraic] GF? More informally, do these objects exhibit a structure similar to the structure of regular [context-free] languages? Is such a structure usually clearly visible? That is to say, is it easy to feel, to predict rationality [algebraicity]?

We shall see that the answer to all these questions tends to be *yes* in the rational case (with a few warnings...) but is probably *no* in the algebraic case. In particular, the rich world of *planar maps* (planar graphs embedded in the sphere) abounds in candidates for non-$\mathbb{N}$-algebraicity. The algebraicity of the associated GFs has been

known for more than 40 years (at least for some families of maps), but it is only in the past 10 years that a general combinatorial explanation of this algebraicity has emerged. Moreover, the underlying constructions are more general that those allowed in context-free descriptions, as they involve taking *complements*.

Each of the main two sections ends with a list of questions. In particular, we present at the end of Section 3 several counting problems that are simple to state and have an algebraic GF, but for reasons that remain mysterious.

The paper is sometimes written in an informal style. We hope that this will not stop the reader. We have tried to give precise references where he/she will find more details and more material on the topics we discuss. In particular, this survey borrows a lot to two books that we warmly recommend: Stanley's *Enumerative Combinatorics* [74], [75], and Flajolet & Sedgewick's *Analytic Combinatorics* [43].

**Notation and definitions.** Given a (commutative) ring $R$, we denote by $R[t]$ the ring of polynomials in $t$ having coefficients in $R$. A *Laurent series* in $t$ is a series of the form $A(t) = \sum_{n \geq n_0} a_n t^n$, with $n_0 \in \mathbb{Z}$ and $a_n \in R$ for all $n$. If $n_0 \geq 0$, we say that $A(t)$ is a *formal power series*. The coefficient of $t^n$ is denoted $a_n := [t^n]A(t)$. The set of Laurent series forms a ring, and even a field if $R$ is a field. The *quasi-inverse* of $A(t)$ is the series $A^*(t) := 1/(1 - A(t))$. If $A(t)$ is a formal power series with constant term 0, then $A^*(t)$ is a formal power series too.

In most occasions, the series we consider are GFs of the form (1) and thus have rational coefficients. However, we sometimes consider refined enumeration problems, in which every object $A$ is *weighted*, usually by a monomial $w(A)$ in some additional indeterminates $x_1, \ldots, x_m$. The weighted GF is then $\sum_{A \in \mathcal{A}} w(A) t^{|A|}$, so that the coefficient ring is $\mathbb{Q}[x_1, \ldots, x_m]$ rather than $\mathbb{Q}$.

We denote $[\![k]\!] = \{1, 2, \ldots, k\}$. We use the standard notation $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{P} := \{1, 2, 3, \ldots\}$.

## 2. Rational generating functions

**2.1. Definitions and properties.** The Laurent series $A(t)$ with coefficients in the field $R$ is said to be *rational* if it can be written in the form

$$A(t) = \frac{P(t)}{Q(t)}$$

where $P(t)$ and $Q(t)$ belong to $R[t]$.

There is probably no need to spend a lot of time explaining why such series are simple and well-behaved. We refer to [74, Ch. 4] and [43, Ch. IV] for a survey of their properties. Let us review briefly some of them, in the case where $R = \mathbb{Q}$. The set of (Laurent) rational series is closed under sum, product, derivation, reciprocals – but *not under integration* as shown by $A(t) = 1/(1 - t)$. The coefficients $a_n$ of a rational series $A(t)$ satisfy a linear recurrence relation with constant coefficients:

for $n$ large enough,

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k}.$$

The partial fraction expansion of $A(t)$ provides a closed form expression of these coefficients of the form:

$$a_n = \sum_{i=0}^{k} P_i(n)\mu_i^n \tag{2}$$

where the $\mu_i$ are the reciprocals of the roots of the denominator $Q(t)$, and the $P_i$ are polynomials. In particular, if $A(t)$ has non-negative integer coefficients, its radius of convergence $\rho$ is one its the poles (Pringsheim) and the "typical" asymptotic behaviour of $a_n$ is

$$a_n \sim \kappa \rho^{-n} n^d \tag{3}$$

where $d \in \mathbb{N}$ and $\kappa$ is an algebraic number. The above statement has to be taken with a grain of salt: *all* poles of minimal modulus may actually contribute to the dominant term in the asymptotic expansion of $a_n$, as indicated by (2).

Let us add that Padé approximants allow us to *guess* whether a generating function whose first coefficients are known is likely to be rational. For instance, given the 10 first coefficients of the series

$$A(t) = t + 2\,t^2 + 6\,t^3 + 19\,t^4 + 61\,t^5 + 196\,t^6 + 629\,t^7 + 2017\,t^8 + 6466\,t^9 + 20727\,t^{10} + O(t^{11}),$$

it is easy to conjecture that actually

$$A(t) = \frac{t\,(1-t)^3}{1 - 5\,t + 7\,t^2 - 4\,t^3}.$$

Padé approximants are implemented in most computer algebra packages. For instance, the relevant MAPLE command is `convert/ratpoly`.

**2.2. Walks on a digraph.** We now introduce our typical "rational" objects. Let $G = (V, E)$ be a directed graph with (finite) vertex set $V = [\![p]\!]$ and (directed) edge set $E \subset V \times V$. A *walk of length* $n$ on $G$ is a sequence of vertices $w = (v_0, v_1, \ldots, v_n)$ such that for all $i$, the pair $(v_i, v_{i+1})$ is an edge. Such a walk *goes from* $v_0$ *to* $v_n$. We denote $|w| = n$. Now assign to each directed edge $e$ a weight (an indeterminate) $x_e$. Define the weight $x_w$ of the walk $w$ as the product of the weights of the edges it visits: more precisely,

$$x_w = \prod_{i=0}^{n-1} x_{(v_i, v_{i+1})}.$$

See Figure 1 (a) for an example. Let $X$ denote the (weighted) adjacency matrix of $G$: for $i$ and $j$ in $[\![p]\!]$, the entry $X_{i,j}$ is $x_e$ if $(i, j) = e$ is an edge of $G$ and 0 otherwise.

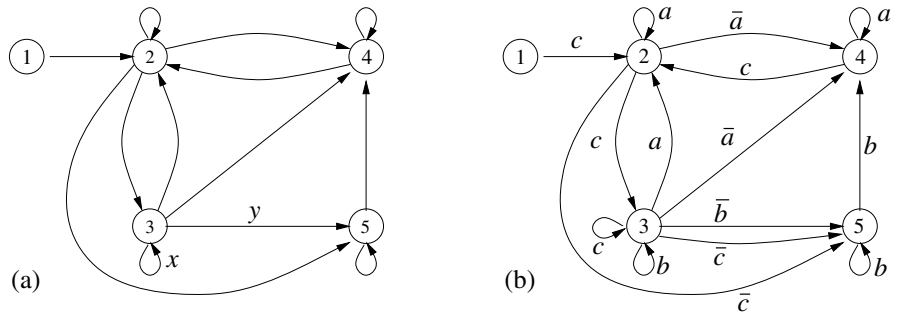



Figure 1. (a) A weighted digraph. The default value of the weight is 1. (b) A deterministic automaton on the alphabet $\{a, b, c, \bar{a}, \bar{b}, \bar{c}\}$. The initial state is 1 and the final states are 2 and 3.

Let $W_{i,j}(t)$ be the weighted generating function of walks going from $i$ to $j$:

$$W_{i,j}(t) = \sum_{w:\, i \rightsquigarrow j} x_w t^{|w|}.$$

It is well-known, and easy to prove, that $W_{i,j}$ is a rational function in $t$ with coefficients in $\mathbb{Q}[x_e, e \in E]$ (see [74, Thm. 4.7.1]).

**Theorem 2.1.** *The series $W_{i,j}(t)$ is the $(i, j)$-entry in the matrix $(1 - tX)^{-1}$.*

This theorem reduces the enumeration of walks on a digraph to the calculation of the inverse of a matrix with polynomial coefficients. It seems to be little known in the combinatorics community that this inverse matrix can be computed by studying the elementary cycles of the digraph $G$. This practical tool relies on Viennot's theory of *heaps of pieces* [81]. Since it is little known, and often convenient, let us advertise it here. It will be illustrated further down.

An *elementary cycle* of $G$ is a closed walk $w = (v_0, v_1, \ldots, v_{n-1}, v_0)$ such that $v_0, \ldots, v_{n-1}$ are distinct. It is defined up to a cyclic permutation of the $v_i$. That is, $(v_1, v_2, \ldots, v_{n-1}, v_0, v_1)$ is the same cycle as $w$. A collection $\gamma = \{\gamma_1, \ldots, \gamma_r\}$ of (elementary) cycles is *non-intersecting* if the $\gamma_i$ are pairwise disjoint. The weight $x_\gamma$ of $\gamma$ is the product of the weights of the $\gamma_i$. We denote $|\gamma| = \sum |\gamma_i|$.

**Proposition 2.2** ([81]). *The generating function of walks going from $i$ to $j$ reads*

$$W_{i,j}(t) = \frac{N_{i,j}}{D},$$

*where*

$$D = \sum_{\gamma = \{\gamma_1, \ldots, \gamma_r\}} (-1)^r x_\gamma t^{|\gamma|} \quad \text{and} \quad N_{i,j} = \sum_{w; \gamma = \{\gamma_1, \ldots, \gamma_r\}} (-1)^r x_w x_\gamma t^{|w|+|\gamma|}.$$

*The polynomial $D$ is the alternating generating function of non-intersecting collections of cycles. In the expression of $N$, $\gamma$ a non-intersecting collection of cycles and $w$ a self-avoiding walk going from $i$ to $j$, disjoint from the cycles of $\gamma$.*

To illustrate this result, let us determine the generating function of walks going from 1 to 2 and from 1 to 3 on the digraph of Figure 1 (a). This graph contains 4 cycles of length 1, 2 cycles of length 2, 2 cycles of length 3 and 1 cycle of length 4. By forming all non-intersecting collections of cycles, one finds:

$$\begin{aligned} D(t) &= 1-(3+x)t+(3+3x-2)t^2+(-1-3x+3+x-2)t^3+(x-1-x+1+x-y)t^4 \\ &= 1-(3+x)t+(1+3x)t^2-2xt^3+(x-y)t^4. \end{aligned}$$

There is only one self-avoiding walk (SAW) going from 1 to 2, and one SAW going from 1 to 3 (via the vertex 2). The collections of cycles that do not intersect these walks are formed of loops, which gives

$$N_{1,2} = t(1-t)^2(1-xt) \quad \text{and} \quad N_{1,3} = t^2(1-t)^2.$$

Hence the generating function of walks that start from 1 and end at 2 or 3 is:

$$W_{1,2}+W_{1,3} = \frac{N_{1,2}+N_{1,3}}{D} = \frac{t(1-t)^2(1+t-xt)}{1-(3+x)t+(1+3x)t^2-2xt^3+(x-y)t^4}. \quad (4)$$

**2.3. Regular languages and automata.** There is a very close connection between the collection of walks on a digraph and the words of *regular languages*. Let $\mathcal{A}$ be an *alphabet*, that is, a finite set of symbols (called *letters*). A *word* on $\mathcal{A}$ is a sequence $u = u_1u_2\dots u_n$ of letters. The number of occurrences of the letter $a$ in the word $u$ is denoted $|u|_a$. The *product* of two words $u_1u_2\dots u_n$ and $v_1v_2\dots v_m$ is the concatenation $u_1u_2\dots u_nv_1v_2\dots v_m$. The empty word is denoted $\varepsilon$. A *language* on $\mathcal{A}$ is a set of words. We define two operations on languages:

- the product $\mathcal{L}\mathcal{K}$ of two languages $\mathcal{L}$ and $\mathcal{K}$ is the set of words $uv$, with $u \in \mathcal{L}$ and $v \in \mathcal{K}$; this product is easily seen to be associative,
- the star $\mathcal{L}^*$ of the language $\mathcal{L}$ is the union of all languages $\mathcal{L}^k$, for $k \geq 0$. By convention, $\mathcal{L}^0$ is reduced to the empty word $\varepsilon$.

A *finite state automaton* on $\mathcal{A}$ is a digraph $(V, E)$ *with possibly multiple edges*, together with:

– a labelling of the edges by letters of $\mathcal{A}$, that is to say, a function $L: E \to \mathcal{A}$,

– an initial vertex $i$,

– a set $V_f \subset V$ of final vertices.

The vertices are usually called the *states* of the automaton. The automaton is *deterministic* if for every state $v$ and every letter $a$, there is at most one edge labelled $a$ starting from $v$.

To every walk on the underlying multigraph, one associates a word on the alphabet $\mathcal{A}$ by reading the letters met along the walk. The language $\mathcal{L}$ *recognized* by the automaton is the set of words associated with walks going from the initial state $i$ to

one of the states of $V_f$. For $j \in V$, let $\mathcal{L}_j$ denote the set of words associated with walks going from $i$ to $j$. These sets admit a recursive description. For the automaton of Figure 1 (b), one has $\mathcal{L} = \mathcal{L}_2 \cup \mathcal{L}_3$ with

$$\begin{aligned}
&\mathcal{L}_1 = \{\varepsilon\},\\
&\mathcal{L}_2 = \mathcal{L}_1 c \cup \mathcal{L}_2 a \cup \mathcal{L}_3 a \cup \mathcal{L}_4 c, \quad \mathcal{L}_4 = \mathcal{L}_2 \bar{a} \cup \mathcal{L}_3 \bar{a} \cup \mathcal{L}_4 a \cup \mathcal{L}_5 b,\\
&\mathcal{L}_3 = \mathcal{L}_2 c \cup \mathcal{L}_3 b \cup \mathcal{L}_3 c, \qquad\qquad \mathcal{L}_5 = \mathcal{L}_2 \bar{c} \cup \mathcal{L}_3 \bar{b} \cup \mathcal{L}_3 \bar{c} \cup \mathcal{L}_5 b.
\end{aligned}$$

Remarkably, there also exists a non-recursive combinatorial description of the languages that are recognized by an automaton [52, Thms. 3.3 and 3.10].

**Theorem 2.3.** *Let $\mathcal{L}$ be a language on the alphabet $\mathcal{A}$. There exists a finite state automaton that recognizes $\mathcal{L}$ if and only if $\mathcal{L}$ can be expressed in terms of finite languages on $\mathcal{A}$, using a finite number of unions, products and stars of languages.*

*If these conditions hold, $\mathcal{L}$ is said to be* regular. *Moreover, there exists a* deterministic *automaton that recognizes $\mathcal{L}$.*

**Regular languages and walks on digraphs.** Take a deterministic automaton, and associate with it a weighted digraph as follows: the vertices are those of the automaton, and for all vertices $j$ and $k$, if $m$ edges go from $j$ to $k$ in the automaton, they are replaced by a *single* edge labelled $m$ in the digraph. For instance, the automaton of Figure 1 (b) gives the digraph to its left, with $x = y = 2$. Clearly, the length GF of words of $\mathcal{L}$ is the GF of (weighted) walks of this digraph going from the initial vertex $i$ to one of the final vertices of $V_f$. For instance, according to (4), the length GF of the language recognized by the automaton of Figure 1 (b) is

$$A(t) = \frac{t\,(1-t)^3}{1 - 5\,t + 7\,t^2 - 4\,t^3}. \tag{5}$$

Take a regular language $\mathcal{L}$ recognized by a deterministic automaton $\mathcal{A}$. There exists another deterministic automaton that recognizes $\mathcal{L}$ and does not contain multiple edges. The key is to create a state $(j, a)$ for every edge labelled $a$ ending at $j$ in the automaton $\mathcal{A}$. The digraph associated with this new automaton has all its edges labelled 1, so that there exists a length preserving bijection between the words of $\mathcal{L}$ and the walks on the digraph going from a specified initial vertex $v_0$ to one of the vertices of a given subset $V_f$ of vertices.

Conversely, starting from a digraph with all edges labelled 1, together with a specified vertex $v_0$ and a set $V_f$ of final vertices, it is easy to construct a regular language that is in bijection with the walks of the graph going from $v_0$ to $V_f$ (consider the automaton obtained by labelling all edges with distinct letters). This shows that *counting words of regular languages is completely equivalent to counting walks in digraphs.* In particular, the set of rational series obtained in both types of problems coincide, and have even been given a name:

**Definition 2.4.** A series $A(t) = \sum_{n\geq 0} a_n t^n$ with coefficients in $\mathbb{N}$ is said to be $\mathbb{N}$-*rational* if there exists a regular language having generating function $A(t) - a_0$.

The description of regular languages given by Theorem 2.3 implies that the set of $\mathbb{N}$-rational series contains the smallest set of series containing $\mathbb{N}[t]$ and closed under sum, product and quasi-inverse. The converse is true [71, Thm. II.5.1]. There exists a simple way to decide whether a given rational series with coefficients in $\mathbb{N}$ is $\mathbb{N}$-rational [71, Thms. II.10.2 and II.10.5].

**Theorem 2.5.** *A series $A(t) = \sum_{n\geq 0} a_n t^n$ with coefficients in $\mathbb{N}$ is $\mathbb{N}$-rational if and only if there exists a positive integer $p$ such that for all $r \in \{0, \dots, p\}$, the series*

$$A_{r,p}(t) := \sum_{n\geq 0} a_{np+r} t^n$$

*has a unique singularity of minimal modulus* (*called* dominant).

There exist rational series with non-negative integer coefficients that are *not* $\mathbb{N}$-rational. For instance, let $\alpha$ be such that $\cos\alpha = 3/5$ and $\sin\alpha = 4/5$, and define $a_n = 25^n \cos(n\alpha)^2$. It is not hard to see that $a_n$ is a non-negative integer. The associated series $A(t)$ reads

$$A(t) = \frac{1 - 2t + 225t^2}{(1 - 25t)(625t^2 + 14t + 1)}.$$

It has 3 distinct dominant poles. As $\alpha$ is not a rational multiple of $\pi$, the same holds for all series $A_{0,p}(t)$, for all values of $p$. Thus $A(t)$ is not $\mathbb{N}$-rational.

**2.4. The combinatorial intuition of rational generating functions.** We have described two families of combinatorial objects that naturally yield rational generating functions: walks in a digraph and words of regular languages. We have, moreover, shown that the enumeration of these objects are equivalent problems. It seems that these families convey the "right" intuition about objects with a rational GF. By this, we mean informally that:

(i) "every" family of objects with a rational GF has actually an $\mathbb{N}$-rational GF,

(ii) for almost all families of combinatorial objects with a rational GF, it is easy to foresee that there will be a bijection between these objects and words of a regular language.

Point (ii) means that most of these families $\mathcal{F}$ have a clear *automatic structure*, similar to the automatic structure of regular languages: roughly speaking, the objects of $\mathcal{F}$ can be constructed recursively using unions of sets and concatenation of *cells* (replacing letters). A more formal definition would simply paraphrase the definition of automata.

Point (i) means simply that I have never met a counting problem that would yield a rational, but not $\mathbb{N}$-rational GF. This includes problems coming from algebra, like growth functions of groups. On the contrary, Point (ii) only concerns purely combinatorial problems (but I do not want to be asked about the border between

combinatorics and algebra). It admits very few counter-examples. Some will be discussed in Section 2.5. For the moment, let us illustrate the two above statements by describing the automatic structure of certain classes of objects (some being rather general), borrowed from [74, Ch. 4].

**2.4.1. Column-convex polyominoes.** A *polyomino* is a finite union of cells of the square lattice, whose interior is connected. Polyominoes are considered up to a translation. A polyomino is *column-convex* (cc) if its intersection with every vertical line is connected. Let $a_n$ be the number of cc-polyominoes having $n$ cells, and let $A(t)$ be the associated generating function. We claim that these polyominoes have an automatic structure.

Consider a cc-polyomino $P$ having $n$ cells. Let us number these cells from 1 to $n$ as illustrated in Figure 2. The columns are visited from left to right. In the first column, cells are numbered from bottom to top. In each of the other columns, the lowest cell that has a left neighbour gets the smallest number; then the cells lying

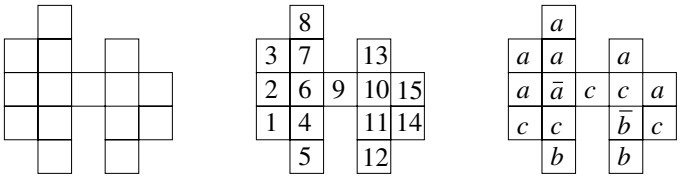



Figure 2. A column-convex polyomino, with the numbering and encoding of the cells.

below it are numbered from top to bottom, and finally the cells lying above it are numbered from bottom to top. Note that for all $i$, the cells labelled $1, 2, \ldots, i$ form a cc-polyomino. This will be essential in our description of the automatic structure of these objects. Associate with $P$ the word $u = u_1 \ldots u_n$ on the alphabet $\{a, b, c\}$ defined by

– $u_i = c$ (like Column) if the $i$th cell is the first to be visited in its column,

– $u_i = b$ (like Below) if the $i$th cell lies below the first visited cell of its column,

– $u_i = a$ (like Above) if the $i$th cell lies above the first visited cell of its column.

Then, add a bar on the letter $u_i$ if the $i$th cell of $P$ has a South neighbour, an East neighbour, but no South-East neighbour. (In other words, the barred letters indicate where to start a new column, when the bottommost cell of this new column lies above the bottommost cell of the previous column.) This gives a word $v$ on the alphabet $\{a, b, c, \bar{a}, \bar{b}, \bar{c}\}$. It is not hard to see that the map that sends $P$ on the word $v$ is a size-preserving bijection between cc-polyominoes and words recognized by the automaton of Figure 1 (b). Hence by (5), the generating function of column-convex polyominoes is [76]:

$$A(t) = \frac{t\,(1-t)^3}{1 - 5\,t + 7\,t^2 - 4\,t^3}.$$

**2.4.2. *P*-partitions.** A *partition* of the integer $n$ into at most $k$ parts is a non-decreasing $k$-tuple $\lambda = (\lambda_1, \dots, \lambda_k)$ of nonnegative integers that sum to $n$. This classical number-theoretic notion is generalized by the notion of *P-partitions*. Let $P$ be a *natural* partial order on $[\![k]\!]$ (by *natural* we mean that if $i < j$ in $P$, then $i < j$ in $\mathbb{N}$). A *$P$-partition* of $n$ is a $k$-tuple $\lambda = (\lambda_1, \dots, \lambda_k)$ of nonnegative integers that sum to $n$ and satisfy $\lambda_i \leq \lambda_j$ if $i \leq j$ in $P$. Thus when $P$ is the natural total order on $[\![k]\!]$, a $P$-partition is simply a partition[1].

We are interested in the following series:

$$F_P(t) = \sum_{\lambda} t^{|\lambda|},$$

where the sum runs over all $P$-partitions and $|\lambda| = \lambda_1 + \cdots + \lambda_k$ is the *weight* of $\lambda$.

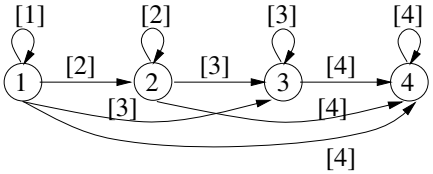
The case of ordinary partitions is easy to analyze: every partition can be written in a unique way as a linear combination

$$c_1\lambda^{(1)} + \cdots + c_k\lambda^{(k)} \tag{6}$$

where $\lambda^{(i)} = (0, 0, \dots, 0, 1, 1, \dots, 1)$ has exactly $i$ parts equal to 1 and $c_i \in \mathbb{N}$. The weight of $\lambda^{(i)}$ is $i$, and one obtains:

$$F_P(t) = \frac{1}{(1-t)(1-t^2)\dots(1-t^k)}. \tag{7}$$

The automatic structure of (ordinary) partitions is transparent: since they are constructed by adding a number of copies of $\lambda^{(1)}$, then a number of copies of $\lambda^{(2)}$, and so on, there is a size preserving bijection between these partitions and walks starting from 1 and ending anywhere in the following digraph:




Note that this graph corresponds to $k = 4$, and that an edge labelled $[\ell]$ must be understood as a *sequence* of $\ell$ edges. These labels do *not* correspond to multiplicities. Observe that the only cycles in this digraph are loops. This, combined with Proposition 2.2, explains the factored form of the denominator of (7).

Consider now the partial order on $[\![4]\!]$ defined by $1 < 3$, $2 < 3$ and $2 < 4$. The partitions of weight at most 2 are

$$(0,0,0,0),\ (0,0,1,0),\ (0,0,0,1),\ (1,0,1,0),\ (0,0,1,1),\ (0,0,2,0),\ (0,0,0,2),$$

[1] A $P$-partition is usually defined as an *order-reversing* map from $[\![k]\!]$ to $\mathbb{N}$ [74, Section 4.5]. Both notions are of course completely equivalent.

so that $F_P(t) = 1 + 2t + 4t^2 + O(t^3)$. If one is brave enough to list $P$-partitions of weight at most 20, the Padé approximant of the truncated series thus obtained is remarkably simple:

$$F_P(t) = \frac{1 + t + t^2 + t^3 + t^4}{(1-t)(1-t^2)(1-t^3)(1-t^4)} + O(t^{21}),$$

and allows one to make a (correct) conjecture.

It turns out that the generating function of $P$-partitions is always a rational series of denominator $(1-t)(1-t^2)\dots(1-t^k)$. Moreover, $P$-partitions obey our general intuition about objects with a rational GF. The following proposition, illustrated below by an example, describes their automatic structure: the set of $P$-partitions can be partitioned into a finite number of subsets; in each of these subsets, partitions have a structure similar to (6). Recall that a *linear extension* of $P$ is a bijection $\sigma$ on $[\![k]\!]$ such that $\sigma(i) < \sigma(j)$ if $i <_P j$.

**Proposition 2.6** ([74], Section 4.5). *Let $P$ be a natural order on $[\![k]\!]$.*

*For every $P$-partition $\lambda$, there exists a unique linear extension $\sigma$ of $P$ such that for all $i$, $\lambda_{\sigma(i)} \le \lambda_{\sigma(i+1)}$, the inequality being strict if $\sigma(i) > \sigma(i+1)$. We say that $\lambda$ is* compatible *with $\sigma$.*

*Given a linear extension $\sigma$, the $P$-partitions that are compatible with $\sigma$ can be written in a unique way as a linear combination with coefficients in $\mathbb{N}$:*

$$\lambda^{(\sigma,0)} + c_1\lambda^{(\sigma,1)} + \cdots + c_k\lambda^{(\sigma,k)} \tag{8}$$

*where $\lambda^{(\sigma,0)}$ is the smallest $P$-partition compatible with $\sigma$:*

$$\lambda^{(\sigma,0)}_{\sigma(j)} = |\{i < j : \sigma(i) > \sigma(i+1)\}| \quad \text{for } 1 \le j \le k,$$

*and for $1 \le i \le k$,*

$$(\lambda^{(\sigma,i)}_{\sigma(1)}, \dots, \lambda^{(\sigma,i)}_{\sigma(k)}) = (0, 0, \dots, 0, 1, 1, \dots, 1)$$

*has exactly $i$ parts equal to 1. Thus the GF of these $P$-partitions is*

$$F_{P,\sigma}(t) = \frac{t^{e(\sigma)}}{(1-t)(1-t^2)\dots(1-t^k)}$$

*where $e(\sigma)$ is a variant of the* Major index *of $\sigma$:*

$$e(\sigma) = \sum_{i:\sigma(i)>\sigma(i+1)} (k-i).$$

**Example.** Let us return to the order $1 < 3$, $2 < 3$ and $2 < 4$. The 5 linear extensions are 1234, 2134, 1243, 2143 and 2413. Take $\sigma = 2143$. The $P$-partitions $\lambda$ that are compatible with $\sigma$ are those that satisfy $\lambda_2 < \lambda_1 \le \lambda_4 < \lambda_3$. The smallest of them is thus $\lambda^{(\sigma,0)} = (1, 0, 2, 1)$. Then $\lambda^{(\sigma,1)} = (0, 0, 1, 0)$, $\lambda^{(\sigma,2)} = (0, 0, 1, 1)$, $\lambda^{(\sigma,3)} = (1, 0, 1, 1)$ and $\lambda^{(\sigma,4)} = (1, 1, 1, 1)$.

**2.4.3. Integer points in a convex polyhedral cone ([74], Sec. 4.6).** Let $\mathcal{H}$ be a finite collection of linear half-spaces of $\mathbb{R}^m$ of the form $c_1\alpha_1 + \cdots + c_m\alpha_m \geq 0$, with $c_i \in \mathbb{Z}$. We are interested in the set $\mathcal{E}$ of *non-negative integer points* $\alpha = (\alpha_1, \ldots, \alpha_m)$ lying in the intersection of those half-spaces. For instance, we could have the following set $\mathcal{E}$, illustrated in Figure 3 (a):

$$\mathcal{E} = \{(\alpha_1, \alpha_2) \in \mathbb{N}^2 : 2\alpha_1 \geq \alpha_2 \text{ and } 2\alpha_2 \geq \alpha_1\}. \tag{9}$$

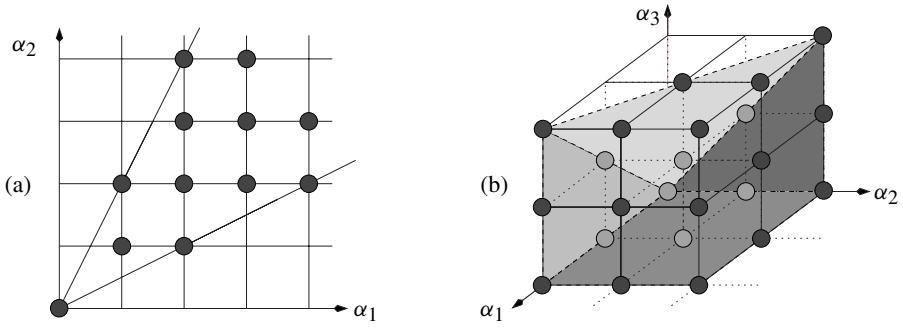



Figure 3. Integer points in a polyhedral cone.

Numerous enumerative problems (including $P$-partitions) can be formulated in terms of linear inequalities as above. The generating function of $\mathcal{E}$ is

$$E(t) = \sum_{\alpha \in \mathcal{E}} t^{|\alpha|},$$

where $|\alpha| = \alpha_1 + \cdots + \alpha_m$. In the above example, $E(t) = 1 + t^2 + 2t^3 + t^4 + 2t^5 + 3t^6 + 2t^7 + O(t^8)$.

The set $\mathcal{E}$ is a monoid (it is closed under summation). In general, it is not a *free* monoid. Geometrically, the set $\mathcal{C}$ of non-negative *real* points in the intersection of the half-spaces of $\mathcal{H}$ forms a *pointed convex polyhedral cone* (the term *pointed* means that it does not contain a line), and $\mathcal{E}$ is the set of integer points in $\mathcal{C}$.

**The simplicial case.** In the simplest case, the cone $\mathcal{C}$ is simplicial. This implies that the *monoid* $\mathcal{E}$ is *simplicial*, meaning that there exists linearly independent vectors $\alpha^{(1)}, \ldots, \alpha^{(k)}$ such that

$$\mathcal{E} = \{\alpha \in \mathbb{N}^m : \alpha = q_1\alpha^{(1)} + \cdots + q_k\alpha^{(k)} \text{ with } q_i \in \mathbb{Q},\ q_i \geq 0\}.$$

This is the case in Example (9), with $\alpha^{(1)} = (1, 2)$ and $\alpha^{(2)} = (2, 1)$. The *interior* of $\mathcal{E}$ (the set of points of $\mathcal{E}$ that are not on the boundary of $\mathcal{C}$) is then

$$\overline{\mathcal{E}} = \{\alpha \in \mathbb{N}^m : \alpha = q_1\alpha^{(1)} + \cdots + q_k\alpha^{(k)} \text{ with } q_i \in \mathbb{Q}, q_i > 0\}. \tag{10}$$

Then there exists a finite subset $\mathcal{D}$ of $\mathcal{E}$ [resp. $\overline{\mathcal{D}}$ of $\overline{\mathcal{E}}$] such that every element of $\mathcal{E}$ [resp. $\overline{\mathcal{E}}$] can be written uniquely in the form

$$\alpha = \beta + c_1\alpha^{(1)} + \cdots + c_k\alpha^{(k)}, \tag{11}$$

with $\beta \in \mathcal{D}$ [resp. $\beta \in \overline{\mathcal{D}}$] and $c_i \in \mathbb{N}$ [74, Lemma 4.6.7]. In our running example (9), taken with $\alpha^{(1)} = (1, 2)$ and $\alpha^{(2)} = (2, 1)$, one has $\mathcal{D} = \{(0, 0), (1, 1), (2, 2)\}$ while $\overline{\mathcal{D}} = \{(1, 1), (2, 2), (3, 3)\}$. Compare (11) with the structure found for $P$-partitions (8). Thus $\mathcal{E}$ and $\overline{\mathcal{E}}$ have an automatic structure and their GFs read

$$E(t) = \frac{\sum_{\beta\in\mathcal{D}} t^{|\beta|}}{\prod_{i=1}^{k}(1 - t^{|\alpha^{(i)}|})} \quad \text{resp.} \quad \overline{E}(t) = \frac{\sum_{\beta\in\overline{\mathcal{D}}} t^{|\beta|}}{\prod_{i=1}^{k}(1 - t^{|\alpha^{(i)}|})}.$$

In Example (9), one thus obtains

$$E(t) = \frac{1 + t^2 + t^4}{(1 - t^3)^2} = \frac{1 - t + t^2}{(1 - t)(1 - t^3)} \quad \text{and} \quad \overline{E}(t) = t^2E(t).$$

**The general case.** The set $\mathcal{E}$ can always be partitioned into a finite number of sets $\overline{\mathcal{F}}$ of the form (10), where $\mathcal{F}$ is a simplicial monoid [74, Ch. 4, Eq. (24)]. Thus $\mathcal{E}$, as a finite union of sets with an automatic structure, has an automatic structure as well. The associated generating function $E(t)$ is $\mathbb{N}$-rational, with a denominator which is a product of cyclotomic polynomials.

Consider, for example, the set

$$\mathcal{E} = \{(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^3 : \alpha_3 \le \alpha_1 + \alpha_2\}.$$

The cone $\mathcal{C}$ of non-negative *real* points $\alpha$ satisfying $\alpha_3 \le \alpha_1+\alpha_2$ is *not* simplicial, as it has 4 faces of dimension 2, lying respectively in the hyperplanes $\alpha_i = 0$ for $i = 1, 2, 3$ and $\alpha_3 = \alpha_1+\alpha_2$ (Figure 3 (b)). But it is the union of two simplicial cones $\mathcal{C}_1$ and $\mathcal{C}_2$, obtained by intersecting $\mathcal{C}$ with the half-spaces $\alpha_1 \ge \alpha_3$ and $\alpha_1 \le \alpha_3$, respectively. Let $\mathcal{E}_1$ [resp. $\mathcal{E}_2$] denote the set of integer points of $\mathcal{C}_1$ [resp. $\mathcal{C}_2$].

The fastest way to obtain the generating function $E(t)$ is to write

$$E(t) = E_1(t) + E_2(t) - E_{12}(t) \tag{12}$$

where $E_{12}(t)$ counts integer points in the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$ (that is, in the plane $\alpha_1 = \alpha_3$). Since $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_1 \cap \mathcal{E}_2$ are simplicial cones (of dimension 3, 3 and 2 respectively), the method presented above for simplicial cones applies. Indeed, $\mathcal{E}_1$ [resp. $\mathcal{E}_2$; $\mathcal{E}_{12}$] is the set of linear combinations (with coefficients in $\mathbb{N}$) of $(1, 0, 1)$, $(0, 1, 0)$ and $(1, 0, 0)$ [resp. $(1, 0, 1)$, $(0, 1, 0)$ and $(0, 1, 1)$; $(1, 0, 1)$ and $(0, 1, 0)$]. This implies:

$$E(t) = \frac{1}{(1 - t)^2(1 - t^2)} + \frac{1}{(1 - t)(1 - t^2)^2} - \frac{1}{(1 - t)(1 - t^2)} = \frac{1 + t + t^2}{(1 - t)(1 - t^2)^2}.$$

However, the "minus" sign in (12) prevents us from seeing directly the automatic nature of $\mathcal{E}$ (the difference of $\mathbb{N}$-rational series is not always $\mathbb{N}$-rational). This structure only becomes clear when we write $\mathcal{E}$ as the disjoint union of the interiors of all simplicial monoids induced by the triangulation of $\mathcal{C}$ into $\mathcal{C}_1$ and $\mathcal{C}_2$. These monoids are the integer points of the faces (of all possible dimensions) of $\mathcal{C}_1$ and $\mathcal{C}_2$. As there are 12 such faces (more precisely, 1 [resp. 4, 5, 2] faces of dimension 0 [resp. 1, 2, 3]), this gives $\mathcal{E}$ as the disjoint union of 12 sets having an automatic structure of the form (10).

### 2.5. Rational generating functions: more difficult questions.

**2.5.1. Predicting rationality.** We wrote in Section 2.4 that it is usually easy to foresee, to predict when a class of combinatorial objects has a rational GF. There are a few exceptions. Here is one of the most remarkable ones.

**Example 2.7** (*Directed animals*). A *directed animal with a compact source of size k* is a finite set of points $A$ on the square lattice $\mathbb{Z}^2$ such that:

– the points $(-i, i)$ for $0 \leq i < k$ belong to $A$; they are called the *source points*,

– all the other points in $A$ can be reached from one of the source points by a path made of North and East steps, having all its vertices in $A$.

See Figure 4 for an illustration. A similar notion exists for the triangular lattice. It turns out that these animals have extremely simple generating functions [50], [10].

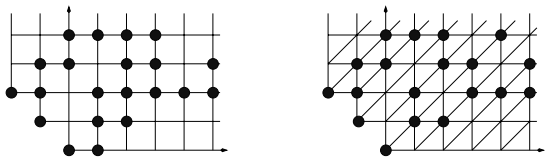


Figure 4. Compact-source directed animals on the square and triangular lattices.

**Theorem 2.8.** *The number of compact-source directed animals of cardinality $n$ is $3^{n-1}$ on the square lattice, and $4^{n-1}$ on the triangular lattice.*

The corresponding GFs are respectively $t/(1-3t)$ and $t/(1-4t)$, and are as rational as a series can be. There is at the moment no simple combinatorial intuition as to why these animals have rational GFs. A bijection between square lattice animals and words on a 3-letter alphabet was described in [50], but it does not shed a clear light on the structure of these objects. Still, there is now a convincing explanation of the *algebraicity* of these series (see Section 3.4.2).

**Example 2.9** (*The area under Dyck paths*). Another family of (slightly less natural) examples is provided by the enumeration of points lying below certain lattice paths.

For instance, let us call *Dyck path of length* $2n$ any path $P$ on $\mathbb{Z}^2$ formed of steps $(1, 1)$ and $(1, -1)$, that starts from $(0, 0)$ and ends at $(2n, 0)$ without ever hitting a point with a negative ordinate. The *area* below $P$ is the number of non-negative integer points $(i, j)$, with $i \leq 2n$, lying weakly below $P$ (Figure 5). It turns out that the sum of the areas of Dyck paths of length $2n$ is simply

$$\sum_{P:|P|=2n} a(P) = 4^n.$$

Again, the rationality of the associated generating function does not seem easy to predict, but there are good combinatorial reasons explaining why it is *algebraic*. See [33], [65] for a direct explanation of this result, references, and a few variations on this phenomenon, first spotted by Kreweras [58].

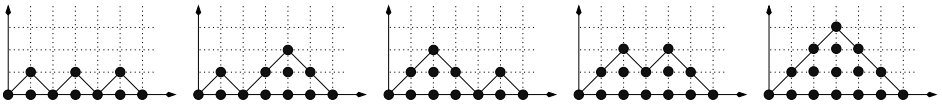


Figure 5. The 5 Dyck paths of length 6 and the $4^3 = 64$ points lying below.

Finally, let us mention that our optimistic statement about how easy it is to predict the rationality of a generating function becomes less and less true as we move from purely combinatorial problems to more algebraic ones. For instance, it is not especially easy to foresee that a group has an automatic structure [39]. Let us give also an example coming from number theory. Let $P(x) \equiv P(x_1, \ldots, x_r)$ be a polynomial with integer coefficients, and take $p$ a prime. For $n \geq 0$, let $a_n$ be the number of $x \in (\mathbb{Z}/p^n\mathbb{Z})^r$ such that $P(x) = 0 \mod p^n$. Then the generating function $\sum_n a_n t^n$ is rational. A related result holds with $p$-adic solutions [37], [53].

**2.5.2. Computing a rational generating function.** Let us start with an elementary, but important observation. Many enumerative problems, including some very hard, can be *approximated* by problems having a rational GF. To take one example, consider the notoriously difficult problem of counting *self-avoiding polygons* (elementary cycles) on the square lattice. It is easy to convince oneself that the generating function of SAP lying in a horizontal strip of height $k$ is rational for all $k$. This does not mean that it will be easy (or even possible, in the current state of affairs) to compute the corresponding generating function when $k = 100$. Needless to say, there is at the moment no hope to express this GF for a *generic* value of $k$. The generating function of SAP having $2k$ horizontal steps can also be seen to be rational. Moreover, these SAP can be described in terms of linear inequalities (as in Section 2.4.3), which implies that the denominator of the corresponding series $G_k$ is a product of cyclotomic polynomials. But again, no one knows what this series is for a generic value of $k$, or even for $k = 100$. Still, some progress have been made recently, since it has been

proved that the series $G_k$ have more and more poles as $k$ increases, which means that their denominators involve infinitely many cyclotomic polynomials [68]. This may be considered as a *proof of the difficulty* of this enumerative problem [51].

In general, computing the (rational) generating function of a family of objects depending on a parameter $k$ may be non-obvious, if not difficult, even if the objects are clearly regular, and even if the final result turns out to be nice. A classical example is provided by the growth functions of Coxeter groups [61]. Here is a more combinatorial example. A partition $\lambda = (\lambda_1, \ldots, \lambda_k)$ is said to be a *$k$-Lecture Hall partition* ($k$-LHP) if

$$0 \leq \frac{\lambda_1}{1} \leq \frac{\lambda_2}{2} \leq \cdots \leq \frac{\lambda_k}{k}.$$

Since these partitions are defined by linear inequalities, it follows from Section 2.4.3 that their weight generating function is rational, with a denominator formed of cyclotomic polynomials. Still, there is no clear reason to expect that [15]:

$$\sum_{\lambda\ k-\mathrm{LHP}} t^{|\lambda|} = \frac{1}{(1-t)(1-t^3)\ldots(1-t^{2k-1})}.$$

Several proofs have been given for this result and variations on it. See for instance [16], [35] and references in the latter paper. Some of these proofs are based on a bijection between lecture hall partitions and partitions into parts taken in $\{1, 3, \ldots, 2k-1\}$, but these bijections are never really simple [82], [40].

**2.5.3. $\mathbb{N}$-rationality.** As we wrote in Section 2.4, we do not know of a counting problem that would yield a rational, but not $\mathbb{N}$-rational series. It would certainly be interesting to find one (even if it ruins some parts of this paper).

Let us return to Soittola's criterion for $\mathbb{N}$-rationality (Theorem 2.5). It is not always easy to prove that a rational series has non-negative coefficients. For instance, it was conjectured in [46] that for any odd $k$, the number of partitions of $n$ into parts taken in $\{k, k+1, \ldots, 2k-1\}$ is a non-decreasing function of $n$, for $n \geq 1$. In terms of generating functions, this means that the series

$$q + \frac{1-q}{(1-q^k)(1-q^{k+1})\ldots(1-q^{2k-1})}$$

has non-negative coefficients. This was only proved recently [67]. When $k$ is even, a similar result holds for the series

$$q + \frac{1-q}{(1-q^k)(1-q^{k+1})\ldots(1-q^{2k})(1-q^{2k+1})}.$$

Once the non-negativity of the coefficients has been established, it is not hard to prove that these series are $\mathbb{N}$-rational. This raises the question of finding a family of combinatorial objects that they count.

## 3. Algebraic generating functions

**3.1. Definitions and properties.** The Laurent series $A(t)$ with coefficients in the field $R$ is said to be *algebraic* (over $R(t)$) if it satisfies a non-trivial *algebraic* equation:

$$P(t, A(t)) = 0$$

where $P$ is a bivariate polynomial with coefficients in $R$. We assume below $R = \mathbb{Q}$.

Again, the set of algebraic Laurent series possesses numerous interesting properties [75, Ch. 6], [43, Ch. VII]. It is closed under sum, product, derivation, reciprocals, but not under integration. These closure properties become effective using either the theory of elimination or Gröbner bases, which are implemented in most computer algebra packages. The coefficients $a_n$ of an algebraic series $A(t)$ satisfy a linear recurrence relation with polynomial coefficients: for $n$ large enough,

$$p_0(n)a_n + p_1(n)a_{n-1} + p_2(n)a_{n-2} + \cdots + p_k(n)a_{n-k} = 0.$$

Thus the first $n$ coefficients can be computed using a linear number of operations.

There is no systematic way to express the coefficients of an algebraic series in closed form. Still, one can sometimes apply the *Lagrange inversion formula*:

**Proposition 3.1.** *Let $\Phi$ and $\Psi$ be two formal power series and let $U \equiv U(t)$ be the unique formal power series with no constant term satisfying*

$$U = t\Phi(U).$$

*Then for $n > 0$, the coefficient of $t^n$ in $\Psi(U)$ is:*

$$[t^n]\Psi(U) = \frac{1}{n}[t^{n-1}]\left(\Psi'(t)\Phi(t)^n\right).$$

Given an algebraic equation $P(t, A(t)) = 0$, one can decide whether there exists a series $U(t)$ and two *rational* series $\Phi$ and $\Psi$ satisfying

$$U = t\Phi(U) \quad \text{and} \quad A = \Psi(U). \tag{13}$$

Indeed, such series exist if and only if the *genus* of the curve $P(t, a)$ is zero [1, Ch. 15]. Moreover, both the genus and a parametrization of the curve in the form (13) can be determined algorithmically.

**Example 3.2 (Finding a rational parametrization).** The following algebraic equation was recently obtained [22], after a highly non-combinatorial derivation, for the GF of certain planar graphs carrying a *hard-particle configuration*:

$$\begin{aligned} 0 = {} & 23328\,t^6A^4 + 27\,t^4(91 - 2088\,t)A^3 \\ & + t^2(86 - 3951\,t + 46710\,t^2 + 3456\,t^3)A^2 \\ & + (1 - 69\,t + 1598\,t^2 - 11743\,t^3 - 14544\,t^4)A \\ & - 1 + 66\,t - 1495\,t^2 + 11485\,t^3 + 128\,t^4. \end{aligned} \tag{14}$$

The package `algcurves` of MAPLE, and more precisely the commands `genus` and `parametrization`, reveal that a rational parametrization is obtained by setting

$$t = -3\,\frac{(3\,U+7)\left(9\,U^2+33\,U+37\right)}{(3\,U+1)^4}.$$

Of course, this is just the net result of MAPLE, which is not necessarily very meaningful for combinatorics. Still, starting from this parametrization, one obtains after a few attempts an alternative parametrizing series $V$ with positive coefficients:

$$V = \frac{t}{(1-2V)(1-3V+3V^2)}. \tag{15}$$

The main interest of such a parametrization for this problem does *not* lie in the possibility of applying the Lagrange inversion formula. Rather, it suggests that a more combinatorial approach exists, based on the enumeration of certain *trees*, in the vein of [19], [27]. It also gives a hint of what these trees may look like.

Another convenient tool borrowed from the theory of algebraic curves is the possibility to explore all branches of the curve $P(t, A(t)) = 0$ in the neighbourhood of a given point $t_0$. This is based on Newton's polygon method. All branches have a *Puiseux expansion*, that is, an expansion of the form:

$$A(t) = \sum_{n \geq n_0} a_n (t - t_0)^{n/d}$$

with $n_0 \in \mathbb{Z}$, $d \in \mathbb{P}$. The coefficients $a_n$ belong to $\mathbb{C}$ (in general, to an algebraic closure of the ground field). These expansions can be computed automatically using standard software. For instance, the MAPLE command `puiseux` of the `algcurves` package tells us that (14) has a unique solution that is a formal power series, the other three solutions starting with a term $t^{-2}$.

Such Puiseux expansions are crucial for studying the *asymptotic behaviour* of the coefficients of an algebraic series $A(t)$. As in the rational case, one has first to locate the singularities of $A(t)$, considered as a function of a complex variable $t$. These singularities are found among the roots of the *discriminant* and of the leading coefficient of $P(t, a)$ (seen as a polynomial in $a$). The singular expansion of $A(t)$ near its singularities of smallest modulus can then be converted, using certain *transfer theorems*, into an asymptotic expansion of the coefficients [42], [43, VII.4].

**Example 3.3** (*Asymptotics of the coefficients of an algebraic series*). Consider the series $V(t)$ defined by (15). Its singularities lie among the roots of the discriminant

$$\Delta(t) = -3 + 114t - 4635t^2 + 55296t^3.$$

Only one root is real. Denote it $t_0 \sim 0.065$. The modulus of the other two roots is smaller than $t_0$, so they could, in theory, be candidates for singularities. However,

$V(t)$ has non-negative coefficients, and this implies, by Pringsheim's theorem, that one of the roots of minimal modulus is real and positive. Hence $V(t)$ has a unique singularity, lying at $t_0$. A Puiseux expansion at this point gives

$$V(t) = c_0 - c_1\sqrt{1 - t/t_0} + O(t - t_0),$$

for some explicit (positive) algebraic numbers $c_0$ and $c_1$, which translates into

$$[t^n]V(t) = \frac{c_1}{2\sqrt{\pi}} t_0^{-n} n^{-3/2} \left(1 + o(1)\right).$$

The determination of asymptotic expansions for the coefficients of algebraic series is probably not far from being completely automated, at least in the case of series with non-negative coefficients [31], [43]. The "typical" behaviour is

$$a_n \sim \frac{\kappa}{\Gamma(d+1)} \rho^{-n} n^d, \tag{16}$$

where $\kappa$ is an algebraic number and $d \in \mathbb{Q} \setminus \{-1, -2, -3, \dots\}$. Compare with the result (3) obtained for rational series. Again, the above statement is not exact, as the contribution of *all* dominant singularities must be taken into account. See [43, Thm. VII.6] for a complete statement.

Let us add that, again, one can guess if a series $A(t)$ given by its first coefficients satisfies an algebraic equation $P(t, A(t)) = 0$ of a given bi-degree $(d, e)$. The guessing procedure requires to know at least $(d + 1)(e + 1)$ coefficients, and amounts to solving a system of linear equations. It is implemented in the package `Gfun` of MAPLE [72]. For instance, given the 10 first coefficients of the series $V(t)$ satisfying $V(0) = 0$ and (15), one automatically conjectures (15).

**3.2. Plane trees.** Our typical "algebraic" objects will be (plane) trees. Let us begin with their usual intuitive recursive definition. A *tree* is a graph formed of a distinguished vertex (called the *root*) to which are attached a certain number (possibly zero) of trees, ordered from left to right. The number of these trees is the *degree* of the root. The roots of these trees are the *children* of the root. A more rigorous definition describes a tree as a finite set of words on the alphabet $\mathbb{P}$ satisfying certain conditions [63]. We hope that our less formal definition and Figure 6 (a) suffice to understand what we mean. The vertices of a tree are often called *nodes*. Nodes of degree 0 are called *leaves*, the others are called *inner* nodes.

The enumeration of classes of trees yields very often algebraic equations. Let us consider for instance the *complete binary trees*, that is, the trees in which all vertices have degree 0 or 2 (Figure 12). Let $a_n$ be the number of such trees having $n$ leaves. Then, by looking at the two (sub)trees of the root, one gets, for $n > 1$:

$$a_n = \sum_{k=1}^{n-1} a_k a_{n-k}.$$

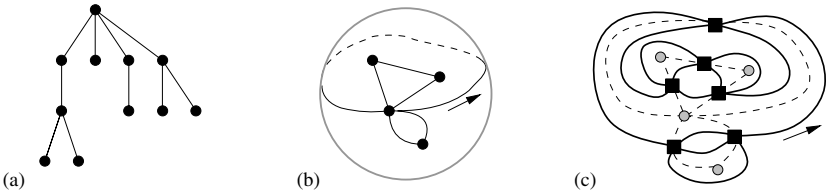



Figure 6. (a) A plane tree. (b) A rooted planar map. (c) The corresponding 4-valent map (thick lines).

The initial condition is $a_1 = 1$. In terms of GFs, this gives $A(t) = t + A(t)^2$, which is easily solved:

$$A(t) = \frac{1-\sqrt{1-4t}}{2} = \sum_{n\geq 0} \frac{1}{n+1}\binom{2n}{n} t^{n+1}. \tag{17}$$

More generally, many algebraic series obtained in enumeration are given as the first component of the solution of a system of the form

$$A_i = P_i(t, A_1, \ldots, A_k), \tag{18}$$

for some polynomials $P_i(t, x_1, \ldots, x_k)$ having coefficients in $\mathbb{Z}$. This system is said to be *proper* if $P_i$ has no constant term ($P_i(0, \ldots, 0) = 0$) and does not contain any linear term $x_i$. It is *positive* if the coefficients of the $P_i$ are non-negative. For instance,

$$A_1 = t^2 + A_1 A_2 \quad \text{and} \quad A_2 = 2tA_1^3$$

is a proper positive system. The system is *quadratic* if every $P_i(t, x_1, \ldots, x_k)$ is a linear combination of the monomials $t$ and $x_\ell x_m$, for $1 \leq \ell \leq m \leq k$.

**Theorem 3.4** ([75], Thm. 6.6.10 and [71], Thm. IV.2.2). *A proper algebraic system has a unique solution $(A_1, \ldots, A_k)$ in the set of formal power series in $t$ with no constant term. This solution is called the* canonical *solution of the system. The series $A_1$ is also the first component of the solution of*

*– a proper quadratic system,*

*– a proper system of the form $B_i = tQ_i(t, B_1, \ldots, B_\ell)$, for $1 \leq i \leq \ell$.*

*These two systems can be chosen to be positive if the original system is positive.*

*Proof.* Let us prove the last property, which we have not found in the above references. Assume $A_1$ satisfies (18) and that this system is quadratic. The $i$th equation reads $A_i = m_i t + n_i A_{\sigma(i)} A_{\tau(i)}$. Rewrite each monomial $A_i A_j$ as $tU_{ij}$ and add the equations $U_{ij} = t\left(m_i m_j + m_i n_j U_{\sigma(j)\tau(j)} + m_j n_i U_{\sigma(i)\tau(i)} + n_i n_j U_{\sigma(i)\tau(i)} U_{\sigma(j)\tau(j)}\right)$. The new system has the required properties. □

**Definition 3.5.** A series $A(t)$ is $\mathbb{N}$-*algebraic* if it has coefficients in $\mathbb{N}$ and if $A(t) - A(0)$ is the first component of the solution of a proper positive system.

Proper positive systems like (18) can always be given a combinatorial interpretation in terms of trees. Every vertex of these trees carries a label $(i, c)$ where $i \in [\![k]\!]$ and $c \in \mathbb{P}$. We say that $i$ is the *type* of the vertex and that $c$ is its colour. The *type* of a tree is the type of its root. Write $A_0 = t$, so that $A_i = P_i(A_0, A_1, \ldots, A_k)$. Let $\mathcal{A}_0$ be the set reduced to the tree with one node, labelled $(0, 1)$. For $i \in [\![k]\!]$, let $\mathcal{A}_i$ be the set of trees such that

– the root has type $i$,

– the types of the subtrees of the root, visited from left to right, are $0, \ldots, 0, 1, \ldots, 1, \ldots, k, \ldots, k$, in this order,

– if exactly $e_j$ children of the root have type $j$, the colour of the root is any integer in the interval $[1, m]$, where $m$ is the coefficient of $x_0^{e_0} \ldots x_k^{e_k}$ in $P_i(x_0, \ldots, x_k)$.

Then it is not hard to see that $A_i(t)$ is the generating function of trees of type $i$, counted by the number of leaves. This explains why trees will be, in the rest of this paper, our typical "algebraic" objects.

**3.3. Context-free languages.** As in the case of rational (and, more precisely, $\mathbb{N}$-rational) series, there exists a family of languages that is closely related to algebraic series. A *context-free grammar* $G$ consists of

– a set $\mathcal{S} = \{S_1, \ldots, S_k\}$ of *symbols*, with one distinguished symbol, say, $S_1$,

– a finite alphabet $\mathcal{A}$ of *letters*, disjoint from $\mathcal{S}$,

– a set of *rewriting rules* of the form $S_i \to w$ where $w$ is a non-empty word on the alphabet $\mathcal{S} \cup \mathcal{A}$.

The grammar is *proper* if there is no rule $S_i \to S_j$. The language $\mathcal{L}(G)$ generated by $G$ is the set of words *on the alphabet* $\mathcal{A}$ that can be obtained from $S_1$ by applying iteratively the rewriting rules. A language is *context-free* is there exists a context-free grammar that generates it. In this case there exists also a proper context-free grammar that generates it.

**Example 3.6** (*Dyck words*). Consider the grammar $G$ having only one symbol, $S$, alphabet $\{a, b\}$, and rules $S \to ab + abS + aSb + aSbS$ (which is short for $S \to ab$, $S \to abS$, $S \to aSb$, $S \to aSbS$). It is easy to see that $\mathcal{L}(G)$ is the set of non-empty words $u$ on $\{a, b\}$ such that $|u|_a = |u|_b$ and for every prefix $v$ of $u$, $|v|_a \geq |v|_b$. These words, called *Dyck words*, provide a simple encoding of the Dyck paths met in Example 2.9.

A *derivation tree* associated with $G$ is a plane tree in which all inner nodes are labelled by symbols, and all leaves by letters, in such a way that if a node is labelled $S_i$ and its children $w_1, \ldots, w_k$ (from left to right), then the rewriting rule $S_i \to w_1 \ldots w_k$ is in the grammar. If the root is labelled $S_1$, then the word obtained by reading the labels of the leaves in prefix order (*i.e.*, from left to right) belongs to the language generated by $G$. Conversely, for every word $w$ in $\mathcal{L}(G)$, there exists at least one derivation tree with root labelled $S_1$ that gives $w$. The grammar is said to be *unambiguous* if every word of $\mathcal{L}(G)$ admits a unique derivation tree.

Assume $G$ is proper. For $1 \leq i \leq k$, let $A_i(t)$ be the generating function of derivation trees rooted at $S_i$, counted by the number of leaves. With each rule $r$, associate the monomial $M(r) = x_0^{e_0} \dots x_k^{e_k}$ where $e_0$ [resp. $e_i$, with $i > 0$] is the number of letters of $\mathcal{A}$ [resp. occurrences of $S_i$] in the right-hand side of $r$. Then the series $A_1, \dots, A_k$ form the canonical solution of the proper positive system (18), with

$$P_i(x_0, x_1, \dots, x_k) = \sum_r M(r),$$

where the sum runs over all rules $r$ with left-hand side $S_i$.

Conversely, starting from a positive system $B_i = tQ_i(t, B_1, \dots, B_k)$ and its canonical solution, it is always possible to construct an unambiguous grammar with symbols $S_1, \dots, S_k$ such that $B_i$ is the generating function of derivation trees rooted at $S_i$ (the idea is to introduce a new letter $a_i$ for each occurrence of $t$). In view of Theorem 3.4 and Definition 3.5, this gives the following alternative characterization of $\mathbb{N}$-algebraic series:

**Proposition 3.7.** *A series $A(t)$ is $\mathbb{N}$-algebraic if and only if only $A(0) \in \mathbb{N}$ and there exists an unambiguous context-free language having generating function $A(t) - A(0)$.*

**3.4. The combinatorial intuition of algebraic generating functions.** We have described two families of combinatorial objects that naturally yield algebraic GFs: plane trees and words of unambiguous context-free languages. We have, moreover, shown a close relationship between these two types of objects. These two families convey the standard intuition of what a family with an algebraic generating function looks like: the algebraicity suggests that it *may* (or should...) be possible to give a recursive description of the objects based on disjoint union of sets and concatenation of objects. Underlying such a description is a context-free grammar. This intuition is the basis of the so-called Schützenberger methodology, according to which the "right" combinatorial way of proving algebraicity is to describe a bijection between the objects one counts and the words of an unambiguous context-free language. This approach has led in the 80s and 90s to numerous satisfactory explanations of the algebraicity of certain series, and we describe some of them in this subsection. Let us, however, warn the reader that the similarities with the rational case will stop here. Indeed, it seems that the "context-free" intuition is far from explaining all algebraicity phenomena in enumerative combinatorics. In particular,

(i) it is very likely that many families of objects have an algebraic, but not $\mathbb{N}$-algebraic generating function,

(ii) there are many families of combinatorial objects with an algebraic GF that do not exhibit a clear "context-free" structure, based on union and concatenation. For several of these families, there is just no explanation of this type, be it clear or not.

This will be discussed in the next subsections. For the moment, let us illustrate the "context-free" intuition.

**3.4.1. Walks on a line.** Let $\mathcal{S}$ be a finite subset of $\mathbb{Z}$. Let $\mathcal{W}$ be the set of walks on the line $\mathbb{Z}$ that start from 0 and take their steps in $\mathcal{S}$. The *length* of a walk is its number of steps. Let $\mathcal{W}_k$ be the set of walks ending at position $k$. For $k \geq 0$, let $\mathcal{M}_k$ be the subset of $\mathcal{W}_k$ consisting of walks that never visit a negative position, and let $\mathcal{M}$ be the union of the sets $\mathcal{M}_k$. In probabilistic terms, the walks in $\mathcal{M}$ would be called *meanders* and the walks of $\mathcal{M}_0$ *excursions.* Of course, a walk is simply a sequence of steps, hence a word on the alphabet $\mathcal{S}$. Thus the sets of walks we have defined can be considered as languages on this alphabet.

**Theorem 3.8.** *The language $\mathcal{W}$ is simply $\mathcal{S}^*$ and is thus regular. The languages $\mathcal{M}$, $\mathcal{W}_k$ and $\mathcal{M}_k$ are unambiguous context-free for all $k$.*

*Proof.* We only describe the (very simple) case $\mathcal{S} = \{+1, -1\}$, to illustrate the ideas that are involved in the construction of the grammar. We encode the steps $+1$ by the letter $a$, the steps $-1$ by $b$, and introduce some auxiliary languages:

- $\mathcal{M}_0^-$, the subset of $\mathcal{W}_0$ formed of walks that never visit a positive position,
- $\mathcal{W}_0^+$ [resp. $\mathcal{W}_0^-$], the subset of $\mathcal{W}_0$ formed of walks that start with $a$ [resp. $b$].

The language $\mathcal{M}_0$ will be generated from the symbol $M_0$, and similarly for the other languages. By looking at the first time a walk of $\mathcal{M}_0$ [resp. $\mathcal{M}_0^-$] reaches position 0 after its first step, one obtains

$$M_0 \to a(1 + M_0)b(1 + M_0) \quad \text{and} \quad M_0^- \to b(1 + M_0^-)a(1 + M_0^-).$$

By considering the last visit to 0 of a walk of $\mathcal{M}_k$, one obtains, for $k > 0$:

$$M_k \to (1 + M_0)a\,(1_{k=1} + M_{k-1})\,.$$

This is easily adapted to general meanders:

$$M \to M_0 + (1 + M_0)a\,(1 + M)\,.$$

Considering the first step of a walk of $\mathcal{W}_0$ gives

$$W_0 \to W_0^+ + W_0^- \quad \text{with} \quad W_0^+ \to M_0(1 + W_0^-) \text{ and } W_0^- \to M_0^-(1 + W_0^+).$$

Finally, for $k > 0$, looking at the first visit at 1 [resp. $-1$] of a walk of $\mathcal{W}_k$ [resp. $\mathcal{W}_{-k}$] yields

$$W_k \to (1 + M_0^-)a\,(1_{k=1} + W_{k-1}) \quad [\text{resp. } W_{-k} \to (1 + M_0)b(1_{k=1} + W_{-(k-1)})].$$

For a general set of steps $\mathcal{S}$, various grammars have been described for the languages $\mathcal{M}_k$ of meanders [38], [60], [59]. For $\mathcal{W}_k$, we refer to [59, Section 4] where the (representative) case $\mathcal{S} = \{-2, -1, 0, 1, 2\}$ is treated. □

Theorem 3.8 is often described in terms of walks in $\mathbb{Z}^2$ starting from $(0, 0)$ and taking their steps in $\{(1, j),\ j \in \mathcal{S}\}$. The conditions on the positions of the walks

that lead to the definition of $\mathcal{M}_k$ and $\mathcal{W}_k$ are restated in terms of conditions on the *ordinates* of the vertices visited by the walk. A harmless generalization is obtained by taking steps in a finite subset $\mathcal{S}$ of $\mathbb{P} \times \mathbb{Z}$. A walk is still encoded by a word on the alphabet $\mathcal{S}$. The languages $\mathcal{W}_k$ remain unambiguous context-free. If each step $(i, j)$ is, moreover, weighted by a rational number $w_{i,j}$, then the generating function of walks of $\mathcal{W}$, counted by the coordinates of their endpoint, is

$$W(t, s) = \frac{1}{1 - \sum_{(i,j)\in\mathcal{S}} w_{i,j} t^i s^j}.$$

The generating function $W_k(t)$ that counts (weighted) walks ending at ordinate $k$ is the coefficient of $s^k$ in $W(t, s)$. Since $\mathcal{W}_k$ is unambiguous context-free, the series $W_k(t)$ is algebraic. This gives a combinatorial explanation of the following result [75, Thm. 6.3.3].

**Theorem 3.9** (Diagonals of rational series). *Let $A(x, y) = \sum_{m,n\geq 0} a_{m,n} x^m y^n$ be a series in two variables $x$ and $y$, with coefficients in $\mathbb{Q}$, that is rational. Then the diagonal of $A$, that is, the series $\Delta A(t) = \sum_{n\geq 0} a_{n,n} t^n$, is algebraic.*

*Proof.* By linearity, it suffices to consider the case

$$A(x, y) = \frac{x^a y^b}{1 - \sum_{0\leq m,n\leq d} c_{m,n} x^m y^n},$$

with $c_{0,0} = 0$. Set $x = ts$ and $y = t/s$. The diagonal of $A$ satisfies

$$\Delta A(t^2) = [s^0]A(ts, t/s) = t^{a+b}[s^{b-a}]\frac{1}{1 - \sum_{0\leq m,n\leq d} c_{m,n} t^{m+n} s^{m-n}},$$

which is algebraic as it counts weighted paths in $\mathcal{W}_{b-a}$, for a certain set of steps. Hence $\Delta A(t)$ is algebraic too. □

The converse of Theorem 3.9 holds: every series $B(t)$ that is algebraic over $\mathbb{Q}(t)$ is the diagonal of a bivariate rational series $A(x, t)$ [70].

**Note.** If one is simply interested in obtaining a set of algebraic equations defining the GFs of the sets $\mathcal{M}_k$ and $\mathcal{W}_k$, a more straightforward approach is to use a partial fraction decomposition (for $\mathcal{W}_k$) and the kernel method (for $\mathcal{M}_k$). See [75, 6.3], and [17, Example 3].

**3.4.2. Directed animals.** Let us move to an example where a neat context-free exists, but is uneasy to discover. We return to the *directed animals* defined in Section 2.5.1. As discussed there, there is no simple explanation as to why the number of compact-source animals is so simple (Theorem 2.8). Still, there is a convincing explanation for the *algebraicity* of the corresponding series: directed animals have, indeed, a context-free structure. This structure was discovered a few years after the proof of

Theorem 2.8, with the development by Viennot of the theory of *heaps* [81], a geometric version of partially commutative monoids [30]. Intuitively, a heap is obtained by dropping vertically some solid pieces, the one after the other. Thus, a piece lies either on the "floor" (then it is said to be *minimal*), or covers, at least partially, another piece.

Directed animals *are*, in essence, heaps. To see this, replace every point of the animal by a *dimer* (Figure 7). Note that if the animal has a unique source, the associated heap has a unique minimal piece. Such heaps are named *pyramids*.

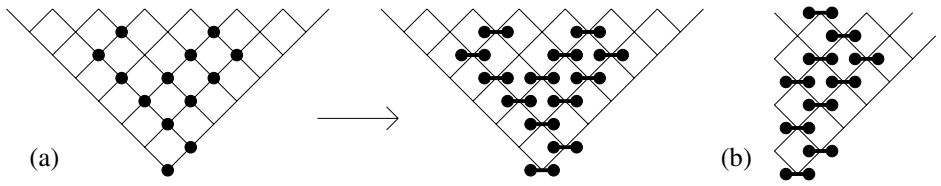



Figure 7. (a) A directed animal and the associated pyramid. (b) A half-pyramid.

What makes heaps interesting here is that there exists a *monoid structure* on the set of heaps: The product of two heaps is obtained by putting one heap above the other and dropping its pieces. This product is the key in our context-free description of directed animals.

Let us begin with the description of pyramids (one-source animals). A pyramid is either a *half-pyramid* (Figure 7 (b)), or the product of a half-pyramid and a pyramid (Figure 8, top). Let $P(t)$ denote the GF of pyramids counted by the number of dimers, and $H(t)$ denote the GF of half-pyramids. Then $P(t) = H(t)(1 + P(t))$. Now, a half-pyramid may be reduced to a single dimer. If it has several dimers, it is the product of a single dimer and of one or two half-pyramids (Figure 8, bottom), which implies $H(t) = t + tH(t) + tH^2(t)$.

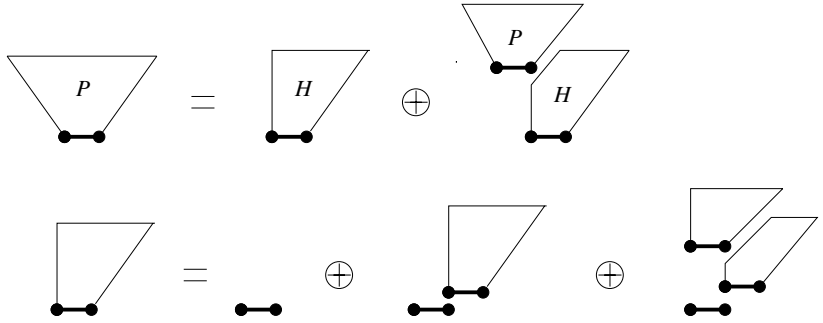



Figure 8. Decomposition of pyramids (top) and half-pyramids (bottom).

A trivial computation finally provides the GF of directed (single-source) animals:

$$P(t) = \frac{1}{2}\left(\sqrt{\frac{1+t}{1-3t}} - 1\right) \qquad \left(\text{while } H(t) = \frac{1-t-\sqrt{(1+t)(1-3t)}}{2t}\right).$$

The enumeration of compact-source directed animals is equivalent to the enumeration of heaps having a compact *basis* (the minimal dimers are adjacent). The generating function of heaps having a compact basis formed with $k$ dimers is $P(t)H(t)^{k-1}$ (Figure 9), which implies that the generating function of compact-source animals is

$$\frac{P(t)}{1-H(t)} = \frac{t}{1-3t}.$$

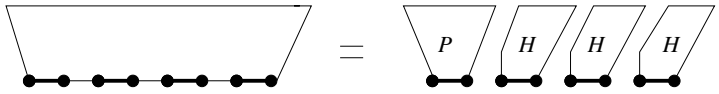



Figure 9. Decomposition of heaps having a compact basis.

**3.5. The world of planar maps.** We have seen in Section 3.2 that plane trees are the paradigm for objects with an algebraic generating function. A more general family of plane objects seems to be just as deeply associated with algebraic series, but for reasons that are far more mysterious: *planar maps*.

A (planar) map is a proper embedding of a planar graph in the sphere (Figure 6 (b)). In order to avoid symmetries, all the maps we consider are *rooted*: this means that one edge is distinguished and oriented. Maps are only considered up to a continuous deformation of the sphere. A map induces a 2-cell decomposition of the sphere: the cells of dimension 0 [resp. 1, 2] are called *vertices* [resp. *edges*, *faces*]. Hence plane trees are maps with a single face.

The interest for the enumeration of planar maps dates back to the early 60s, in connection with the 4-colour theorem. The first results are due to Tutte [77], [78], [79]. Ten to fifteen years later, maps started to be investigated independently in theoretical physics, as a model for 2-dimensional *quantum gravity* [28], [9]. However, neither the recursive approach used by Tutte and his disciples, nor the physics approach based on matrix integrals were able to explain in a combinatorially satisfactory way the following observations:

– the generating functions of many classes of planar maps are algebraic,

– the associated numbers are often irritatingly simple.

Let us illustrate this with three examples.

**1. General maps.** The number of planar maps having $n$ edges is [80]:

$$g_n = \frac{2.3^n}{(n+1)(n+2)}\binom{2n}{n}. \tag{19}$$

The associated generating function $G \equiv G(t) = \sum_{n\geq 0} g_n t^n$ satisfies:

$$-1 + 16t + (1 - 18t)G + 27t^2G^2 = 0. \tag{20}$$

**2. Loopless triangulations.** The number of loopless *triangulations* (maps in which all faces have degree 3) having $2n + 2$ faces is [62]:

$$t_n = \frac{2^n}{(n+1)(2n+1)}\binom{3n}{n}.$$

The associated generating function $T \equiv T(t) = \sum_n t_n t^n$ satisfies

$$1 - 27t + (-1 + 36t)T - 8tT^2 - 16t^2T^3 = 0.$$

**3. Three-connected triangulations.** The number of 3-connected triangulations having $2n + 2$ faces is [77]:

$$m_n = \frac{2}{(n+1)(3n+2)}\binom{4n+1}{n}.$$

The associated generating function $M \equiv M(t) = \sum_n t_n t^n$ satisfies

$$-1 + 16t + (1 - 20t)M + (3t + 8t^2)M^2 + 3t^2M^3 + t^3M^4 = 0.$$

These maps are in bijection with rooted *maximal* planar simple graphs (graphs with no loop nor multiple edge that lose planarity as soon as one adds an edge).

At last, in the past ten years, a general combinatorial picture has emerged, suggesting that maps are, in essence, unrooted plane trees. In what follows, we illustrate on the example of general maps the main three approaches that now exist, and give references for further developments of these methods.

**3.5.1. The recursive approach.** We leave to the reader to experience personally that maps do not have an obvious context-free structure. Still, maps *do* have a simple recursive structure, based on the deletion of the root-edge. However, in order to exploit this structure, one is forced to keep track of the degree of the *root-face* (the face lying to the right of the root edge). The decomposition illustrated in Figure 10 leads in a few lines to the following equation:

$$G(u,t) = 1 + tu^2G(u,t)^2 + tu\,\frac{uG(u,t) - G(1,t)}{u-1}, \tag{21}$$

where $G(u,t)$ counts planar maps by the number of edges ($t$) and the degree of the root-face ($u$).

It can be checked that the above equation defines $G(u,t)$ uniquely as a formal power series in $t$ (with polynomial coefficients in $u$). However, it is not clear on the equation why $G(1,t)$ (and hence $G(u,t)$) are algebraic. In his original paper,

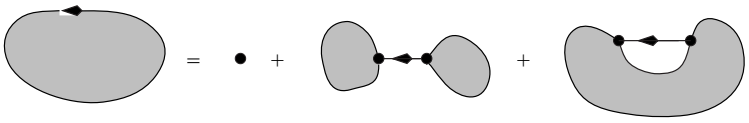


Figure 10. Tutte's decomposition of rooted planar maps.

Tutte first guessed the value of $G_1(t) := G(1,t)$, and then proved the existence of a series $G(u,t)$ that fits with $G_1(t)$ when $u = 1$, and satisfies the above equation. Still, a bit later, Brown came with a *method* for solving (21): the so-called *quadratic method* [29], [49, Sec. 2.9]. Write (21) in the form $(2aG(u,t)+b)^2 = \delta$, where $a$, $b$ and $\delta$ are polynomials in $t$, $u$ and $G_1(t)$. That is,

$$\begin{aligned}&\big(2tu^2(u-1)G(u,t) + tu^2 - u + 1\big)^2 \\ &\qquad = 4t^2u^3(u-1)G_1 + (1-u)^2 - 4tu^4 + 6tu^3 + u^4t^2 - 2tu^2.\end{aligned}$$

It is not hard to see, even without knowing the value of $G(u,t)$, that there exists a (unique) formal power series in $t$, say $U \equiv U(t)$, that cancels the left-hand side of this equation. That is,

$$U = 1 + tU^2 + 2tU^2(U-1)G(U,t).$$

This implies that the series $U$ is a *double root* of the polynomial $\delta$ that lies on the right-hand side. The discriminant of this polynomial (in $u$) thus vanishes: this gives the algebraic equation (20) satisfied by $G(1,t)$.

The enumeration of many other families of planar maps can also be attacked by a recursive description based on the deletion of an edge (or vertex, or face...). See for instance [62] for 2-connected triangulations, or [6] for maps with prescribed face degrees. (For maps with high connectivity, like 3-connected triangulations, an additional *composition formula* is often required [77], [3].) The resulting equations are usually of the form

$$P(F(u), F_1, \ldots, F_k, t, u) = 0, \tag{22}$$

where $F(u) \equiv F(t,u)$, the main generating function, is a series in $t$ with polynomial coefficients in $u$, and $F_1, \ldots, F_k$ are series in $t$ only, independent of $u$. Brown's *quadratic method* applies as long as the degree in $F(u)$ is 2 (for the linear case, see the *kernel method* in [17], [2]). Recently, it was understood how these equations could be solved in full generality [22]. Moreover, the solution of any (well-founded) equation of the above type was shown to be algebraic. This provides two types of enumerative results:

– the proof that many map generating functions are algebraic: it now suffices to exhibit an equation of the form (22), or to explain why such an equation *exists*,

– the solution of previously unsolved map problems (like the enumeration of hard-particle configurations on maps, which led to (14), or that of triangulations with high vertex degrees [8]).

**3.5.2. Matrix integrals.** In the late 70s, it was understood by a group of physicists that certain matrix integral techniques coming from quantum field theory could be used to attack enumerative problems on maps [28], [9]. This approach proved to be extremely efficient (even if it is usually not fully rigorous). The first step is fairly automatized, and consists in converting the description of maps into a certain integral. For instance, the relevant integral for the enumeration of 4-valent maps (maps in which all vertices have degree 4) is

$$Z(t, N) = \frac{2^{N(N-1)/2}}{(2\pi)^{N^2/2}} \int dH e^{\mathrm{tr}(-H^2/2+tH^4/N)},$$

where the integration space is that of hermitian matrices $H$ of size $N$, equipped with the Lebesgue measure $dH = \prod dx_{kk} \prod_{k<\ell} dx_{k\ell} dy_{k\ell}$ with $h_{k\ell} = x_{k\ell} + i y_{k\ell}$. As there is a classical bijection between 4-valent maps with $n$ vertices and planar maps with $n$ edges (Figure 6 (c)), we are still dealing with our reference problem: the enumeration of general planar maps. The connection between the above integral and maps is

$$G(t) = tE'(t) \quad \text{with } E(t) = \lim_{N\to\infty} \frac{1}{N^2} \log Z(t, N).$$

Other map problems lead to integrals involving several hermitian matrices [55]. We refer to [83] for a neat explanation of the encoding of map problems by integrals, and to [45], [41] (and references therein) for the evaluation of integrals.

**3.5.3. Planar maps and trees.** We finally come to a combinatorial explanation of the formula/equation for $g_n$ and $G(t)$. Take a plane binary tree with $n$ (inner) nodes, planted at a leaf, and add to every inner node a new distinguished child, called a *bud*. At each node, we have three choices for the position of the bud (Figure 11 (a)). The new tree, called *budding tree*, has now $n$ buds and $n + 2$ leaves. Now start from the root and walk around the tree in counterclockwise order, paying attention to the sequence of buds and leaves you meet. Each time a bud is immediately followed by

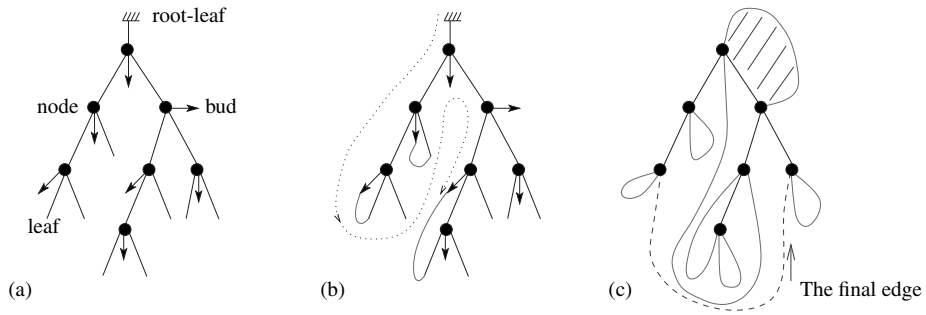



Figure 11. (a). A budding tree. (b) An intermediate step in the matching procedure. (c) The resulting 4-valent map, with its marked face.

a leaf in this sequence, match them by forming a new edge (Figure 11 (b)) and then go on walking around the plane figure thus obtained. At the end, exactly two leaves remain unmatched. Match them together and orient this final edge in one of the two possible ways. Also, mark the face to the left of the matching edge that ends at the root-leaf.

**Theorem 3.10** ([73]). *The above correspondence is a bijection between pairs* $(T, \varepsilon)$ *where* $T$ *is a budding tree having* $n$ *inner nodes and* $\varepsilon \in \{0, 1\}$*, and* 4*-valent maps with* $n$ *vertices and a marked face.*

The value of $\varepsilon$ tells how to orient the final matching edge. Schaeffer first used this bijection to explain combinatorially the formula (19). Indeed, the number of budding trees with $n$ inner nodes is clearly $3^n\binom{2n}{n}/(n+1)$ (see (17)), while the number of 4-valent maps with $n$ vertices and a marked face is $(n+2)g_n$. Eq. (19) follows.

Later, it was realized that this construction could also be used to explain the algebraicity of the series $G(t)$ [23]. Say that a budding tree is *balanced* if the root-leaf is not matched by a bud. Take such a tree, match all buds, and orient the final edge from the root-leaf to the other unmatched leaf. This gives a bijection between balanced budding trees and 4-valent maps. We thus have to count balanced trees, or, equivalently, the unbalanced ones. By re-rooting them at the bud that matches the root-leaf, one sees that they are in bijection with a node attached to three budding trees. This gives

$$G(t) = B(t) - tB(t)^3, \quad \text{where } B(t) = 3t(1 + B(t))^2$$

counts budding trees by (inner) nodes. The above construction involves taking a *difference* of $\mathbb{N}$-algebraic series, which needs not be $\mathbb{N}$-algebraic. We actually conjecture that the series $G(t)$ is not $\mathbb{N}$-algebraic (see Section 3.6.4).

There is little doubt that the above construction (once described in greater detail…) explains in a very satisfactory way both the simplicity of the formula giving $g_n$ and the algebraicity of $G(t)$. Moreover, this is not an *ad hoc*, isolated magic trick: over the past ten years, it was realized that this construction is one in a family of constructions of the same type, which apply to numerous families of maps (Eulerian maps [73], maps with prescribed vertex degrees [23], constellations [18], bipartite maps with prescribed degrees [19], maps with higher connectivity [66], [47]). Definitely, these constructions reveal a lot about the combinatorial nature of planar maps.

To conclude this section, let us mention that a different combinatorial construction for general planar maps, discovered in the early 80s [34], has recently been simplified [32] and adapted to other families of maps [36], [54], [25], [26]. It is a bit less easy to handle than the one based on trees with buds, but it allows one to keep track of the distances between some vertices of the map. This has led to remarkable connections with a random probability distribution called the *Integrated SuperBrownian Excursion* [32]. A third type of construction has emerged even more recently [7] for 2-connected triangulations, but no ones knows at the moment whether it will remain isolated or is just the tip of another iceberg.

**3.6. Algebraic series: some questions.** We begin with three simple classes of objects that have an algebraic GF, but for reasons that remain mysterious. We then discuss a possible criterion (or necessary condition) for $\mathbb{N}$-algebraicity, and finally the algebraicity of certain hypergeometric series.

**3.6.1. Kreweras' words and walks on the quarter plane.** Let $\mathcal{L}$ be the set of words $u$ on the alphabet $\{a, b, c\}$ such that for every prefix $v$ of $u$, $|v|_a \geq |v|_b$ and $|v|_a \geq |v|_c$. These words encode certain walks on the plane: these walks start at $(0, 0)$, are made of three types of steps, $a = (1, 1)$, $b = (-1, 0)$ and $c = (0, -1)$, and never leave the first quadrant of the plane, defined by $x, y \geq 0$. The *pumping lemma* [52, Thm. 4.7], applied to the word $a^n b^n c^n$, shows that the language $\mathcal{L}$ is not context-free. However, its generating function *is* algebraic. More precisely, let us denote by $\ell_{i,j}(n)$ the number of words $u$ of $\mathcal{L}$ of length $n$ such that $|u|_a - |u|_b = i$ and $|u|_a - |u|_c = j$. They correspond to walks of length $n$ ending at position $(i, j)$. Then the associated three-variable generating function is

$$\begin{aligned} L(u, v; t) &= \sum_{i,j,n} \ell_{i,j}(n) u^i v^j t^n \\ &= \frac{\left(1/W - \bar{u}\right)\sqrt{1 - uW^2} + \left(1/W - \bar{v}\right)\sqrt{1 - vW^2}}{uv - t(u + v + u^2v^2)} - \frac{1}{uvt} \end{aligned}$$

where $\bar{u} = 1/u$, $\bar{v} = 1/v$ and $W \equiv W(t)$ is the unique power series in $t$ satisfying $W = t(2 + W^3)$. Moreover, the number of walks ending at $(i, 0)$ is remarkably simple:

$$\ell_{i,0}(3n + 2i) = \frac{4^n(2i + 1)}{(n + i + 1)(2n + 2i + 1)} \binom{2i}{i} \binom{3n + 2i}{n}.$$

The latter formula was proved in 1965 by Kreweras, in a fairly complicated way [57]. This rather mysterious result has attracted the attention of several combinatorialists since its publication [14], [48], [64]. The first combinatorial explanation of the above formula (in the case $i = 0$) has just been found by Bernardi [7].

Walks in the quarter plane do not always have an algebraic GF: for instance, the number of *square lattice walks* (with North, South, East and West steps) of size $2n$ that start and end at $(0, 0)$ and remain in the quarter plane is

$$\frac{1}{(2n + 1)(2n + 4)} \binom{2n + 2}{n + 1}^2 \sim \frac{4^{2n+1}}{\pi n^3},$$

and this asymptotic behaviour prevents the corresponding generating function from being algebraic (see (16)). The above formula is easily proved by looking at the projections of the walk onto the horizontal and vertical axes.

**3.6.2. Walks on the slit plane.** Take now *any* finite set of steps $\mathcal{S} \subset \mathbb{Z} \times \{-1, 0, 1\}$ (we say that these steps have *small height variations*). Let $s_{i,j}(n)$ be the number of

walks of length $n$ that start from the origin, consist of steps of $\mathcal{S}$, never return to the non-positive horizontal axis $\{(-k, 0), k \geq 0\}$, and end at $(i, j)$. Let $S(u, v; t)$ be the associated generating function:

$$S(u, v; t) = \sum_{i,j \in \mathbb{Z}, n \geq 0} s_{i,j}(n) u^i v^j t^n.$$

Then this series is *always* algebraic, as well as the series $S_{i,j}(t) := \sum_n s_{i,j}(n) t^n$ that counts walks ending at $(i, j)$ [13], [20]. For instance, when $\mathcal{S}$ is formed of the usual square lattice steps (North, South, West and East), then

$$S(u, v; t) = \frac{\left(1 - 2t(1 + \bar{u}) + \sqrt{1 - 4t}\right)^{1/2} \left(1 + 2t(1 - \bar{u}) + \sqrt{1 + 4t}\right)^{1/2}}{1 - t(u + \bar{u} + v + \bar{v})}$$

with $\bar{u} = 1/u$ and $\bar{v} = 1/v$. Moreover, the number of walks ending at certain specific points is remarkably simple. For instance:

$$s_{1,0}(2n + 1) = C_{2n+1}, \quad s_{0,1}(2n + 1) = 4^n C_n, \quad s_{-1,1}(2n) = C_{2n},$$

where $C_n = \binom{2n}{n}/(n + 1)$ is the $n$th Catalan number, which counts binary trees (17), Dyck words, and numerous other combinatorial objects [75, Ch. 6]. The first of these three identities has been proved combinatorially [4]. The others still defeat our understanding.

**3.6.3. Embedded binary trees.** We consider again the complete binary trees met at the beginning of Section 3.2. Let us associate with each (inner) node of such a tree a label, equal to the difference between the number of right steps and the number of left steps one takes when going from the root to the node. In other words, the label of the node is its abscissa in the natural integer embedding of the tree (Figure 12).

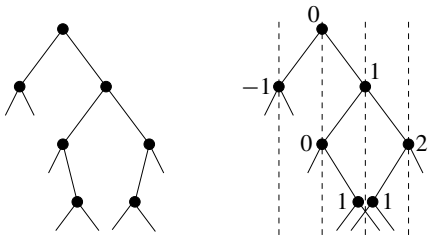



Figure 12. The integer embedding of a binary tree.

Let $S_j \equiv S_j(t, u)$ be the generating function of binary trees counted by the number of nodes (variable $t$) and the number of nodes at abscissa $j$ (variable $u$). Then for all $j \in \mathbb{Z}$, this series is algebraic of degree (at most) 8 (while $S_j(t, 1)$ is quadratic) [12]. Moreover, for $j \geq 0$,

$$S_j = T \, \frac{(1 + \mu Z^j)(1 + \mu Z^{j+5})}{(1 + \mu Z^{j+2})(1 + \mu Z^{j+3})},$$

where

$$T = 1 + tT^2, \qquad Z = t\,\frac{(1+Z^2)^2}{(1-Z+Z^2)},$$

and $\mu \equiv \mu(t,u)$ is the unique formal power series in $t$ satisfying

$$\mu = (u-1)\frac{Z(1+\mu Z)^2(1+\mu Z^2)(1+\mu Z^6)}{(1+Z)^2(1+Z+Z^2)(1-Z)^3(1-\mu^2 Z^5)}.$$

Why is that so? This algebraicity property holds as well for other families of labelled trees [12], [24]. From these series, one can derive certain limit results on the distribution of the number of nodes at abscissa $\lfloor \lambda n^{1/4} \rfloor$ in a random tree with $n$ nodes [12]. These results provide some information about the law of the *integrated super-Brownian excursion* [12], [21].

**3.6.4. $\mathbb{N}$-algebraicity.** $\mathbb{N}$-algebraic series have been defined in Section 3.2 in terms of positive proper algebraic systems. The author has been unable to find in the literature a criterion, or even a necessary condition for an algebraic series with coefficients in $\mathbb{N}$ to be $\mathbb{N}$-algebraic. Nor even an algebraic series with coefficients in $\mathbb{N}$ that would *not* be $\mathbb{N}$-algebraic (together with a proof of this statement...).

A partial answer could be provided by the study of the possible asymptotic behaviour of coefficients of $\mathbb{N}$-algebraic series. It is very likely that not all behaviours of the form (16) are possible. An important result in this direction states that, if a proper positive system (18) is *strongly connected*, the $n$th coefficient of, say, $A_1$ follows the general pattern (16), but with $d = -3/2$ [43, Thm. VII.7]. The system is *strongly connected* if, roughly speaking, the expression of every series $A_i$ involves (possibly after a few iterations of the system) every other series $A_j$. For instance, the system defining the walks ending at 0 in Section 3.4.1 reads

$$M_0 = t^2(1+M_0)^2 \quad \text{and} \quad W_0 = M_0(2+W_0).$$

This system is not strongly connected, as $M_0$ does not involve $W_0$. Accordingly, the number of $2n$-step walks returning to 0 is $\binom{2n}{n} \sim \kappa 4^n n^{-1/2}$.

If one can rule out the possibility that $d = -5/2$ for $\mathbb{N}$-algebraic series, then this will prove that most map generating functions are not $\mathbb{N}$-algebraic (see the examples in Section 3.5).

**3.6.5. Some algebraic hypergeometric series.** Consider the following series:

$$F(t) = \sum_{n\geq 0} f_n t^n = \sum_{n\geq 0} \frac{\prod_{i=1}^{d}(a_i n)!}{\prod_{j=1}^{e}(b_j n)!} t^n,$$

where $a_1, \ldots, a_d, b_1, \ldots, b_e$ are positive integers. This series is algebraic for some values of the $a_i$'s and $b_j$'s, as shown by the case

$$\sum_{n\geq 0} \frac{(2n)!}{n!^2} t^n = \frac{1}{\sqrt{1-4t}}.$$

Can we describe all algebraic cases? Well, one can easily obtain some necessary conditions on the sequences $a$ and $b$ by looking at the asymptotics of $f_n$. First, an algebraic power series has a finite, positive radius of convergence (unless it is a polynomial). This, combined with Stirling's formula, gives at once

$$a_1 + \cdots + a_d = b_1 + \cdots + b_e. \tag{23}$$

Moreover, by looking at the dominant term in the asymptotic behaviour of $f_n$, and comparing with (16), one obtains that either $e = d$, or $e = d+1$. The case $d = e$ only gives the trivial solution $F(t) = 1/(1-t)$, and the complete answer to this problem is as follows [11], [69]:

**Theorem 3.11.** *Assume* (23) *holds and* $F(t) \neq 1/(1-t)$. *The series* $F(t)$ *is algebraic if and only if* $f_n \in \mathbb{N}$ *for all* $n$ *and* $e = d + 1$.

Here are some algebraic instances:

$$f_n = \frac{(6n)!(n)!}{(3n)!(2n)!^2}, \quad f_n = \frac{(10n)!(n)!}{(5n)!(4n)!(2n)!}, \quad f_n = \frac{(20n)!(n)!}{(10n)!(7n)!(4n)!}.$$

The degree of these series is rather big: 12 [resp. 30] for the first [second] series above. This theorem provides a collection of algebraic series with nice integer coefficients: are these series $\mathbb{N}$-algebraic? Do they count some interesting objects?

**Acknowledgements.** The parts of this survey that do not deal exactly with the enumeration of combinatorial objects have often been influenced by discussions with some of my colleagues, including Frédérique Bassino, Henri Cohen, Philippe Flajolet, François Loeser, Géraud Sénizergues. Still, they should not be hold responsible for any of the flaws of this paper.

## References

[1] Abhyankar, S. S., *Algebraic geometry for scientists and engineers*. Math. Surveys Monogr. 35, Amer. Math. Soc., Providence, RI, 1990.

[2] Banderier, C., Bousquet-Mélou, M., Denise, A., Flajolet, P., Gardy, G., and Gouyou-Beauchamps, D., Generating functions for generating trees. *Discrete Math.* **246** (1–3) (2002), 29–55.

[3] Banderier, C., Flajolet, P., Schaeffer, G., and Soria, M., Random maps, coalescing saddles, singularity analysis, and Airy phenomena. *Random Structures Algorithms* **19** (3–4) (2001), 194–246.

[4] Barcucci, E., Pergola, E., Pinzani, R., and Rinaldi, S., A bijection for some paths on the slit plane. *Adv. in Appl. Math.* **26** (2) (2001), 89–96.

[5] Baxter, R. J., *Exactly solved models in statistical mechanics*. Academic Press Inc., London 1982.

[6] Bender, E. A., and Canfield, E. R., The number of degree-restricted rooted maps on the sphere. *SIAM J. Discrete Math.* **7** (1) (1994), 9–15.

[7] Bernardi, O., Bijective counting of Kreweras walks and loopless triangulations. In preparation.

[8] Bernardi, O., On triangulations with high vertex degree. In *Formal Power Series and Algebraic Combinatorics*, Taormina, Italy, 2005; ArXiv math.CO/0601678.

[9] Bessis, D., Itzykson, C., and Zuber, J. B., Quantum field theory techniques in graphical enumeration. *Adv. in Appl. Math.* **1** (2) (1980), 109–157.

[10] Bétréma, J., and Penaud, J.-G., Modèles avec particules dures, animaux dirigés et séries en variables partiellement commutatives. Arxiv:math.CO/0106210.

[11] Beukers, F., and Heckman, G., Monodromy for the hypergeometric function ${}_nF_{n-1}$. *Invent. Math.* **95** (2) (1989), 325–354.

[12] Bousquet-Mélou, M., Limit results for embedded trees. Applications to the integrated super-Brownian excursion. *Random Structures Algorithms*, to appear.

[13] Bousquet-Mélou, M., Walks on the slit plane: other approaches. *Adv. in Appl. Math.* **27** (2–3) (2001), 243–288.

[14] Bousquet-Mélou, M., Walks in the quarter plane: Kreweras' algebraic model. *Ann. Appl. Probab.* **15** (2) (2005), 1451–1491.

[15] Bousquet-Mélou, M., and Eriksson, K., Lecture hall partitions. *Ramanujan J.* **1** (1) (1997), 101–111.

[16] Bousquet-Mélou, M., and Eriksson, K., Lecture hall partitions. II. *Ramanujan J.* **1** (2) (1997), 165–185.

[17] Bousquet-Mélou, M., and Petkovšek, M., Linear recurrences with constant coefficients: the multivariate case. *Discrete Math.* **225** (1–3) (2000), 51–75.

[18] Bousquet-Mélou, M., and Schaeffer, G., Enumeration of planar constellations. *Adv. in Appl. Math.* **24** (4) (2000), 337–368.

[19] Bousquet-Mélou, M., and Schaeffer, G., The degree distribution of bipartite planar maps: applications to the Ising model. ArXiv math.CO/0211070, 2002.

[20] Bousquet-Mélou, M., and Schaeffer, G., Walks on the slit plane. *Probab. Theory Related Fields* **124** (3) (2002), 305–344.

[21] Bousquet-Mélou, M., and Janson, S., The density of the ISE and local limit laws for embedded trees. *Ann. Appl. Probab.*, to appear; ArXiv:math.PR/0509322.

[22] Bousquet-Mélou, M., and Jehanne, A., Planar maps and algebraic series: a generalization of the quadratic method. *J. Combin. Theory Ser. B*, to appear.

[23] Bouttier, J., Di Francesco, P., and Guitter, E., Census of planar maps: from the one-matrix model solution to a combinatorial proof. *Nuclear Phys. B* **645** (3) (2002), 477–499.

[24] Bouttier, J., Di Francesco, P., and Guitter, E., Random trees between two walls: exact partition function. *J. Phys. A* **36** (50) (2003), 12349–12366.

[25] Bouttier, J., Di Francesco, P., and Guitter, E., Statistics of planar graphs viewed from a vertex: a study via labeled trees. *Nuclear Phys. B* **675** (3) (2003), 631–660.

[26] Bouttier, J., Di Francesco, P., and Guitter, E., Planar maps as labeled mobiles. *Electron. J. Combin.* **11** (1) (2004), Research Paper 69, 27 pp. (electronic).

[27] Bouttier, J., Di Francesco, P., and Guitter, E., Combinatorics of bicubic maps with hard particles. *J. Phys. A* **38** (21) (2005), 4529–4559.

[28] Brézin, E., Itzykson, C., Parisi, G., and Zuber, J. B., Planar diagrams. *Comm. Math. Phys.* **59** (1) (1978), 35–51.

[29] Brown, W. G., On the existence of square roots in certain rings of power series. *Math. Ann.* **158** (1965), 82–89.

[30] Cartier, P., and Foata, D., *Problèmes combinatoires de commutation et réarrangements*. Lecture Notes in Math. 85, Springer-Verlag, Berlin 1969.

[31] Chabaud, C., Séries génératrices algébriques ; asymptotique et applications combinatoires. PhD thesis, Université Paris 6, 2002.

[32] Chassaing, P., and Schaeffer, G., Random planar lattices and integrated superBrownian excursion. *Probab. Theory Related Fields* **128** (2) (2004), 161–212.

[33] Chottin, L., and Cori, R., Une preuve combinatoire de la rationalité d'une série génératrice associée aux arbres. *RAIRO Inform. Théor.* **16** (2) (1982), 113–128.

[34] Cori, R., and Vauquelin, B., Planar maps are well labeled trees. *Canad. J. Math.* **33** (5) (1981), 1023–1042.

[35] Corteel, S., Lee, S., and Savage, C. D., Enumeration of sequences constrained by the ratio of consecutive parts. *Sém. Lothar. Combin.* **54** (2005), Art. B54Aa, 12 pp. (electronic).

[36] Del Lungo, A., Del Ristoro, F., and Penaud, J.-G., Left ternary trees and non-separable rooted planar maps. *Theoret. Comput. Sci.* **233** (1–2) (2000), 201–215.

[37] Denef, J., The rationality of the Poincaré series associated to the $p$-adic points on a variety. *Invent. Math.* **77** (1) (1984), 1–23.

[38] Duchon, P., On the enumeration and generation of generalized Dyck words. *Discrete Math.* **225** (1–3) (2000), 121–135.

[39] Epstein, D. B. A., Cannon, J. W., Holt, D. F., Levy, S. V. F., Paterson, M. S., and Thurston, W. P., *Word processing in groups*. Jones and Bartlett Publishers, Boston, MA, 1992.

[40] Eriksen, N., A simple bijection between Lecture Hall partitions and partitions into odd integers. In *Formal Power Series and Algebraic Combinatorics* (Melbourne 2002).

[41] Eynard, B., An introduction to random matrix theory. Technical Report SPHT 01/014, Service de Physique Théorique, CEA Saclay, France, 2001.

[42] Flajolet, P., and Odlyzko, A., Singularity analysis of generating functions. *SIAM J. Discrete Math.* **3** (2) (1990), 216–240.

[43] Flajolet, P., and Sedgewick, R., *Analytic Combinatorics*. Preliminary version available at http://pauillac.inria.fr/algo/flajolet/Publications/books.html.

[44] Flajolet, P., and Sedgewick, R., *An Introduction to the Analysis of Algorithms*. Addison Wesley, Reading, MA, 1996.

[45] Di Francesco, P., 2D Quantum gravity, matrix models and graph combinatorics. 2004; Arxiv:math-ph/0406013.

[46] J. Friedman, J., Joichi, J. T., and Stanton, D., More monotonicity theorems for partitions. *Experiment. Math.* **3** (1) (1994), 31–37.

[47] Fusy, E., Poulalhon, D., and Schaeffer, G., Dissections and trees: applications to optimal mesh encoding and random sampling. In *ACM-SIAM Symposium on Discrete Algorithms*, 2005; http://www.lix. polytechnique.fr/Labo/Gilles.Schaeffer/Biblio/.

[48] Gessel, I. M., A probabilistic method for lattice path enumeration. *J. Statist. Plann. Inference* **14** (1) (1986), 49–58.

[49] Goulden, I. P., and Jackson, D. M., *Combinatorial enumeration*. Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Inc., New York 1983.

[50] Gouyou-Beauchamps, D., and Viennot, G., Equivalence of the two-dimensional directed animal problem to a one-dimensional path problem. *Adv. in Appl. Math.* **9** (3) (1988), 334–357.

[51] Guttmann, A. J., Indicators of solvability for lattice models. *Discrete Math.* **217** (1–3) (2000), 167–189.

[52] Hopcroft, J. E., and Ullman, J. D., *Formal languages and their relation to automata*. Addison Wesley, Reading, MA, 1969.

[53] Igusa, J., Complex powers and asymptotic expansions. II. Asymptotic expansions. *J. Reine Angew. Math.* **278/279** (1975), 307–321.

[54] Jacquard, B., and Schaeffer, G., A bijective census of nonseparable planar maps. *J. Combin. Theory Ser. A* **83** (1) (1998), 1–20.

[55] Kazakov, V. A., Ising model on a dynamical planar random lattice: exact solution. *Phys. Lett. A* **119** (3) (1986), 140–144.

[56] Knuth, D. E., *The art of computer programming*. Vol. 1, 2, 3, Addison Wesley, Reading, MA, 1968–1973.

[57] Kreweras, G., Sur une classe de problèmes liés au treillis des partitions d'entiers. *Cahiers du B.U.R.O.* **6** (1965), 5–105.

[58] Kreweras, G., Aires des chemins surdiagonaux et application à un problème économique. *Cahiers du B.U.R.O.* **24** (1976), 1–8.

[59] Labelle, J., Langages de Dyck généralisés. *Ann. Sci. Math. Québec* **17** (1) (1993), 53–64.

[60] Labelle, J., and Yeh, Y. N., Generalized Dyck paths. *Discrete Math.* **82** (1) (1990), 1–6.

[61] Macdonald, I. G., The Poincaré series of a Coxeter group. *Math. Ann.* **199** (1972), 161–174.

[62] Mullin, R. C., On counting rooted triangular maps. *Canad. J. Math.* **17** (1965), 373–382.

[63] Neveu, J., Arbres et processus de Galton-Watson. *Ann. Inst. H. Poincaré Probab. Statist.* **22** (2) (1986), 199–207.

[64] Niederhausen, H., The ballot problem with three candidates. *European J. Combin.* **4** (2) (1983), 161–167.

[65] Pergola, E., Pinzani, R., Rinaldi, S., and Sulanke, R. A., A bijective approach to the area of generalized Motzkin paths. *Adv. in Appl. Math.* **28** (3–4) (2002), 580–591.

[66] Poulalhon, D., and Schaeffer, G., Optimal coding and sampling of triangulations. In *Automata, languages and programming*, Lecture Notes in Comput. Sci. 2719, Springer, Berlin 2003, 1080–1094.

[67] Prellberg, T., and Stanton, D., Proof of a monotonicity conjecture. *J. Combin. Theory Ser. A* **103** (2) (2003), 377–381.

[68] Rechnitzer, A., Haruspicy 2: The anisotropic generating function of self-avoiding polygons is not D-finite. *J. Combin. Theory Ser. A* **113** (3) (2006), 520–546.

[69] Rodriguez-Villegas, F., Integral ratios of factorials and algebraic hypergeometric functions. Talk given at Oberwolfach, available at http://www.ma.utexas.edu/users/villegas/.

[70] Safonov, K. V., On conditions for the sum of a power series to be algebraic and rational. *Mat. Zametki* **41** (3) (1987), 325–332, 457; English transl. *Math. Notes* **41** (3–4) (19), 185–189.

[71] Salomaa, A., and Soittola, M., *Automata-theoretic aspects of formal power series*. Texts Monogr. Comput. Sci., Springer-Verlag, New York 1978.

[72] Salvy, B., and Zimmermann, P., Gfun: a Maple package for the manipulation of generating and holonomic functions in one variable. *ACM Trans. Math. Software* **20** (2) (1994), 163–177.

[73] Schaeffer, G., Bijective census and random generation of Eulerian planar maps with prescribed vertex degrees. *Electron. J. Combin.* **4** (1) (1997), Research Paper 20, 14 pp. (electronic).

[74] Stanley, R. P., *Enumerative combinatorics*. Vol. 1, Cambridge Stud. Adv. Math. 49, Cambridge University Press, Cambridge 1997.

[75] Stanley, R. P., *Enumerative combinatorics*. Vol. 2, Cambridge Stud. Adv. Math. 62, Cambridge University Press, Cambridge 1999.

[76] Temperley, H. N. V., Combinatorial problems suggested by the statistical mechanics of domains and of rubber-like molecules. *Phys. Rev.* (2) **103** (1956), 1–16 .

[77] Tutte, W. T., A census of planar triangulations. *Canad. J. Math.* **14** (1962), 21–38.

[78] Tutte, W. T., A census of slicings. *Canad. J. Math.* **14** (1962), 708–722.

[79] Tutte, W. T., A census of planar maps. *Canad. J. Math.* **15** (1963), 249–271.

[80] Tutte, W. T., On the enumeration of planar maps. *Bull. Amer. Math. Soc.* **74** (1968), 64–74.

[81] Viennot, G. X., Heaps of pieces. I. Basic definitions and combinatorial lemmas. In *Combinatoire énumérative* (Montréal, 1985), Lecture Notes in Math. 1234, Springer, Berlin 1986, 321–350.

[82] Yee, A. J., On the combinatorics of lecture hall partitions. *Ramanujan J.* **5** (3) (2001), 247–262.

[83] Zvonkin, A., Matrix integrals and map enumeration: an accessible introduction. *Math. Comput. Modelling* **26** (8–10) (1997), 281–304.

CNRS, LaBRI, Université Bordeaux 1, 351 cours de la Libération, 33405 Talence Cedex, France

E-mail: bousquet@labri.fr

# Towards a structure theory for matrices and matroids

Jim Geelen, Bert Gerards, and Geoff Whittle*

**Abstract.** We survey recent work that is aimed at generalizing the results and techniques of the Graph Minors Project of Robertson and Seymour to matrices and matroids.

**Mathematics Subject Classification (2000).** 05B35.

**Keywords.** Matroids, minors, representability, well-quasi-ordering.

## 1. Introduction

We are currently undertaking a program of research aimed at extending the results and techniques of the Graph Minors Project of Robertson and Seymour to matrices and matroids. Here we report on where we stand and where we expect to go.

In particular, we discuss the structure of "minor-closed" classes of matrices over a fixed finite field. This requires a peculiar synthesis of graphs, topology, connectivity, and algebra. In addition to proving several long-standing conjectures in the area, we expect the structure theory will help to find efficient algorithms for a general class of problems on matrices and graphs.

Most combinatorial computational problems are trivial in the sense that they are typically finite. However, even for modest size problems, enumerating the possibilities is practically infeasible; it often results in algorithms whose running time is exponential in the size of the problem. We seek smarter, more efficient, algorithms. In the theory of algorithms *efficient* typically means that the running time is polynomial in the size of the problem.

Often the problems are modeled by graphs (networks) or matrices. The better picture we have of the model, the more likely it is that we can develop a quick algorithm for the problem. For instance, the problem at hand may be more tractable if the modeling graph can be drawn in the plane, or some other particular surface, without crossings. Then it is relevant that we can test efficiently if the graph has such advantageous appearance. Surface embeddability, and other related properties, are preserved when deleting an edge from the graph or contracting an edge (*contracting* means deleting the edge and identifying its ends). The result of any series of such deletions and contractions is called a *minor* of the graph. In this terminology, testing

*This research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada, the Nederlandse Organisatie voor Wetenschappelijk Onderzoek and the Marsden Fund of New Zealand.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

surface embeddability is testing a particular minor-closed graph property. So motivated by real-word computational problems, we end up with the fundamental question if minor-closed graph properties can be tested efficiently.

That this is possible indeed for any fixed minor-closed graph property is one of the consequences of the ground-breaking work of Robertson and Seymour in their Graph Minors Project. One major outcome of this project is their proof of Wagner's Conjecture that graphs are "well-quasi-ordered" under the minor-order, which is the following theorem.

**The Graph Minors Theorem** ([31]). *Any infinite family of graphs contains two members such that one is isomorphic to a minor of the other.*

This implies that for any minor-closed graph property there are only finitely many *excluded minors*, these are graphs that do not have the property but whose proper minors do have the property. For planarity, for instance, there are exactly two excluded minors: $K_5$ and $K_{3,3}$; this is Kuratowski's famous characterization of planarity [22].

So, by the Graph Minors Theorem, to test a minor-closed graph property we only need to test containment of each of its excluded minors individually. That this is possible is another crucial outcome of Robertson and Seymour's work.

**The Graph Minor Recognition Theorem** ([29]). *For each graph $H$, there exists a polynomial-time algorithm for testing if a graph has a minor isomorphic to $H$.*

This answered one of the twelve open problems in Garey and Johnson's 1979 book on computational complexity [8].

So minor-closed graph properties can be tested efficiently. However, as noted earlier, also matrices are widely used as modeling tools, for example in integer programming models for operations research. Integer programming models are very general and powerful, but in a sense too general; they lead to "NP-hard" problems. However if the matrix in an integer programming model is *totally unimodular*, that means if all subdeterminants are 0, 1 or $-1$, then linear programming methods do solve the problem [20], and these methods are efficient. So this raises the issue of testing totally unimodularity, another open problem back in '79, in Garey and Johnson's book. Now, it turns out that a matrix being totally unimodular means that it is in a certain sense representable over any field, and also this embeddability property is closed under certain minor-operations. So also here the fundamental issue of testing minor-closed properties arises. For that we work at extending Robertson and Seymour's graph minor theory to matrices. As the issues involved do not so much concern the actual matrices, but rather the underlying "matroids", we work in that setting.

A *matroid* consists of a finite set $E$, the *ground set* of the matroid, and a function $r$, the *rank function* of the matroid. This rank function is defined on the subsets of $E$ and satisfies the following properties: $0 \leq r(X) \leq |X|$ for $X \subseteq E$; $r(X) \leq r(Y)$ for all $X \subseteq Y \subseteq E$ and $r(X \cup Y) + r(X \cap Y) \leq r(X) + r(Y)$ for all $X, Y \subseteq E$. We call $r(X)$ the *rank* of $X$ and $r(E)$ the *rank* of the matroid. The rank function of a matroid $M$ is

denoted by $r_M$. Two matroids are *isomorphic* if there is a rank-preserving bijection between their ground sets.

Matrices yield matroids: If $A = (a_e : e \in E)$ is a matrix with columns $a_e$ over a field $\mathbb{F}$, then the linear rank of the column submatrices $(a_e : e \in F)$ with $F \subseteq E$ is the rank function of a matroid, the *vector matroid* $M(A)$ of $A$. If a matroid is isomorphic to a vector matroid of a matrix over $\mathbb{F}$, we say that the matroid is *representable over* $\mathbb{F}$ or $\mathbb{F}$-*representable*. (A vector matroid is often described as a configuration of points in a linear, affine or projective space instead of as the collection of columns of a matrix.)

Also graphs yield matroids: Let $G$ be a graph with edge set $E$ and vertex set $V$. The *rank* of a graph is the number of its vertices minus the number of its components. If $F \subseteq E$, then the *rank* of $F$ is the rank of the subgraph of $G$ with edge set equal to $F$. This rank yields the rank function of a matroid, the *cycle matroid* $M(G)$ of $G$. A matroid isomorphic to such a cycle matroid of a graph is called *graphic*. A graphic matroid is representable over any field by a matrix with two non-zero entries in every column, one equal to 1 and one equal to $-1$. We assume the reader to be familiar with the standard notions from graph theory. For matroid theory we refer to Oxley [25] or Welsh [43], but we will define the matroid terminology we use, as we go.

Now we define matroid minors. Let $e$ be an element of the ground set $E$ of a matroid $M$. *Deleting* $e$ *from* $M$ is replacing $M$ by the matroid with ground set $E - \{e\}$ and with rank function equal to the restriction of $r_M$ to subsets of $E - \{e\}$. *Contracting* $e$ *from* $M$ is replacing $M$ by the matroid with ground set $E - \{e\}$ and with rank function $r_M(X \cup \{e\}) - r_M(\{e\})$ for each $X \subseteq E - \{e\}$. A *minor* of a matroid is the result of any sequence of deletions and contractions.

A minor of a vector matroid over a field $\mathbb{F}$ is representable over the field as well. Indeed, deleting an element amounts to just deleting the corresponding column whereas contracting an element $f$ amounts to removing $a_f$ from $(a_e : e \in E)$ and projecting all other columns in the direction of $a_f$ on some arbitrary hyperplane not containing $a_f$. Deletions and contractions in a graph are in one-one correspondence with deletions and contractions in its cycle matroid. Thus the cycle matroid of a minor of a graph is a minor of the cycle matroid of the graph.

So the notion of graph minors is in essence algebraic, or geometric, and in that sense it generalizes to matrices and matroids. This raises the question to what extent Robertson and Seymour's graph minor theory extends to matroids. The following conjecture was by Robertson and Seymour, although to our knowledge not in print.

**The Well-Quasi-Ordering Conjecture.** Let $\mathbb{F}$ be a finite field. Then any infinite set of $\mathbb{F}$-representable matroids contains two matroids, one of which is isomorphic to a minor of the other.

As yet, the Well-Quasi-Ordering Conjecture has not been resolved for any finite field. Note that it is equivalent to the conjecture that, for a finite field $\mathbb{F}$, any minor-closed class of $\mathbb{F}$-representable matroids has a finite number of $\mathbb{F}$-representable excluded minors.

The finiteness of the field in the Well-Quasi-Ordering Conjecture is essential. Indeed, suppose $\mathbb{F}$ is an infinite field and consider for each integer $n \geq 3$, a $2n \times 3$ matrix with columns $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$, where $p_1, \ldots, p_n$ are vectors in general position in $\mathbb{F}^3$ and each $q_i$ is spanned by $p_i$ and $p_{i+1}$, but is not spanned by any other pair among $p_1, \ldots, p_n$ (where $p_{n+1} = p_1$). As $\mathbb{F}$ is infinite, such matrices clearly exist. Among the vector matroids of these matrices none is a minor of another. Indeed, all members of the collection have rank 3, so all minors that use a contraction have too low rank to be in the collection; deleting an element from a member of the collection destroys the unique cyclic arrangement of linearly dependent triples $p_i, q_i, p_{i+1}$ in a way that cannot be repaired by further deletions.

We conjecture additionally that for matroids that are representable over a finite field minor-closed properties can be recognized in polynomial time, in other words we conjecture that also the Graph Minor Recognition Theorem extends.

**The Minor-Recognition Conjecture.** For any finite field $\mathbb{F}$ and any $\mathbb{F}$-representable matroid $N$, there is a polynomial-time algorithm for testing whether an $\mathbb{F}$-representable matroid contains a minor isomorphic to $N$.

At the heart of the Graph Minors Project is Robertson and Seymour's Graph Minors Structure Theorem [30]. It describes constructively the graphs that do not contain a given graph as a minor. This constructive description enables techniques to establish the well-quasi-ordering and algorithmic consequences. For matroids, Seymour [36] used this approach successfully for characterizing total unimodularity (see Section 3). Our hope is to use the same strategy for general matroids that are representable over finite fields. Therefore we are developing a structure theory for such matroids. As a major role in the theory of graph minors is played by connectivity, we need an extension of graph connectivity to matroids.

The basic ingredients of graph connectivity are separations, which tell where the connectivity is not that high, and Menger's Theorem, which provides a way of certifying that the connectivity is not that low. A *separation* of a graph $G$ is a pair $(G_1, G_2)$ of subgraphs of $G$ such that $G = G_1 \cup G_2$; the *order* of the separation $(G_1, G_2)$ is the number of vertices of $G$ that lie in both $G_1$ and $G_2$. A graph is *k-connected* if it has no separation $(G_1, G_2)$ of order $l$ less than $k$ such that $G_1$ and $G_2$ have at least $l$ edges each. One of the fundamental theorems of graph theory is Menger's Theorem.

**Menger's Theorem** ([24]). *If $G$ is a graph and $S$ and $T$ are two sets of vertices, then there either exist $k$ disjoint paths, each connecting a vertex in $S$ with a vertex in $T$, or (exclusively) $G$ has a separation $(G_1, G_2)$ of order less than $k$ such that $S$ lies in $G_1$ and $T$ in $G_2$.*

As an illustration of how this theorem plays a role in finding minors consider the following easy result of Dirac [4]: *A 3-connected graph with at least 4 vertices has a minor isomorphic to $K_4$.* ($K_n$ denotes the complete graph with $n$ vertices; *complete* means that every pair of vertices is connected by an edge.) Here is a proof of Dirac's

result: we may assume the graph has two non-adjacent vertices $s$ and $t$, otherwise the graph is complete and we are done. Apply Menger's Theorem to the set $S$ of neighbours of $s$ and the set $T$ of neighbours of $t$. This yields three $P_1$, $P_2$ and $P_3$ from $s$ to $t$ that only meet at their ends. As $G$ is 3-connected, $G - s - t$ is connected, so there exists a path $Q$ that connects two of these paths and misses the third path. The union of $P_1$, $P_2$, $P_3$ and $Q$ clearly has a minor isomorphic to $K_4$. So Dirac's result follows. This is a very easy result of course, but it may convince the reader of the need of a notion of matroid connectivity and a matroidal version of Menger's Theorem.

A *separation* of a matroid $M$ is a partition $(X, Y)$ of the ground set $E$. The *order* of the separation $(X, Y)$ is $r_M(X) + r_M(Y) - r_M(E) + 1$; for a representable matroid this is the dimension of the intersection of the subspaces spanned by $X$ and $Y$ plus 1. A matroid is *k-connected* if it has no separation $(X, Y)$ of order $l$ less than $k$ such that $X$ and $Y$ have size $l$ or more. If $(G_1, G_2)$ is a separation of $G$ where $E_1$ and $E_2$ are the edge sets of $G_1$ respectively $G_2$, then $(E_1, E_2)$ is a separation of $M(G)$. When $G_1$, $G_2$, and $G$ are connected graphs, the orders of these separations are the same. This is enough to consider matroid separations and matroid connectivity as genuine generalizations of these notions for graphs. And there is a matroidal generalization of Menger's Theorem as well.

**Tutte's Linking Theorem** ([42]). *Let $M$ be a matroid and $X$ and $Y$ be disjoint subsets of its ground set. Then there exists a minor of $M$ in which $(X, Y)$ is a separation of order at least $k$ or (exclusively) $M$ has a separation $(A, B)$ of order less than $k$ with $X$ in $A$ and $Y$ in $B$.*

This follows quite easily from Edmonds' Matroid Intersection Theorem [7], which is one of the fundamental theorems of matroid theory. So we see that the basic theory of graph connectivity does extend quite well to matroids.

Besides the fact that a matroid structure theory will help proving the Well-Quasi-Ordering Conjecture and Minor-Recognition Conjecture for matroids, we also expect that it will provide a handle on the following conjecture, probably the most famous open question in matroid theory.

**Rota's Conjecture** ([33]). Let $\mathbb{F}$ be a finite field. There are, up to isomorphism, only finitely many excluded minors for the class of $\mathbb{F}$-representable matroids.

This also has computational relevance, namely for the question how hard it is to decide if a matroid given by an oracle for the rank function is representable over a field $\mathbb{F}$. Unfortunately, it does take exponentially many oracle calls to decide this, for any field [38]. But if Rota's conjecture is true, then for every finite field $\mathbb{F}$ there exists, for every non-$\mathbb{F}$-representable matroid, a polynomial-length certificate for this non-representability, that requires only a constant number of oracle calls. (It is known that this can be done by a quadratic number of oracle calls [14].)

For the three smallest fields all excluded minors for representability over the field are known: there is one for GF(2) (Tutte [40]), there are four for GF(3) (Bixby [1]

and Seymour [37], independently), and there are seven for GF(4) (Geelen, Gerards and Kapoor [9]). For all other finite fields, Rota's conjecture is still open. A structure theory could well provide a way to prove it.

Also in Rota's conjecture finiteness of the field is essential. Lazarson [23] showed that there are an infinite number of excluded minors for representability over the reals and this is certainly true for all other infinite fields.

Summarizing, en route for these three conjectures we are working at establishing the structure of minor-closed proper subclasses of matroids representable over a finite field. This work has already had some success. It turns out that excluding the cycle matroid of a planar graph as a minor imposes tangible structure on matroids over finite fields, so we begin with discussing that.

## 2. Excluding a planar graph

Let $\mathbb{F}$ be a finite field and let $H$ be a planar graph. We give a constructive structural description of $\mathbb{F}$-representable matroids with no $M(H)$-minor and show that this description enables significant progress on the three conjectures in Section 1. Essentially the structure is to be decomposable into small pieces along low-order separations. We will first explain what that means.

A *branch-decomposition* of matroid $M$ is a tree $T$ in which all vertices have degree 1 or 3, where the degree-1 vertices of $T$ are in 1-1 correspondence with the elements of the ground set $E$ of $M$. The *width* of an edge $e$ in $T$ is the order of the separation $(X, Y)$ of $M$ where $X$ contains the elements of $E$ that correspond to the degree-1 vertices of $T$ in one component of $T - e$ and $Y$ the elements of $E$ that correspond to the degree-1 vertices of $T$ in the other component of $T - e$. So a branch-decomposition is a data-structure for a collection of separations. The *width* of a branch-decomposition is the maximum of the widths of its edges and the *branch-width* of a matroid is the minimum of the widths of all its branch-decompositions. So roughly low branch-width means to be decomposable into small pieces along low-order separations.

Branch-width is a matroid generalization of branch-width for graphs defined by Robertson and Seymour [28]. For graphs it is, up to a constant multiplicative bound, the same as tree-width, also introduced by Robertson and Seymour. In the Graph Minors Project they mainly use tree-width and that notion does extend to matroids [19] as well. But branch-width is easier to work with for matroids, so here we will only use branch-width, also for graphs. Robertson and Seymour prove the following result.

**The Grid Theorem for graphs** ([27]). *For each planar graph $H$ there is an integer $k$ such that any graph with branch-width at least $k$ has a minor isomorphic to $H$.*

This result is called the Grid Theorem because, as every planar graph is a minor of a grid, it suffices to prove it for the case that $H$ is a grid. Here, *a grid*, or rather

*an n by n grid*, refers to the graph with a vertex $(i, j)$ for each pair of integers $i$ and $j$ between 1 and $n$ and an edge between any two pairs $(i_1, j_1)$ and $(i_2, j_2)$ with $|i_1 - i_2| + |j_1 - j_2| = 1$. If a matroid has a minor isomorphic to a cycle matroid of an $n$ by $n$ grid, we say it has a *grid-minor*. To convince oneself that each planar graph $H$ is a minor of a sufficiently large grid, visualize $H$ as drawn without crossings and with the edges and vertices as thick lines and dots on a piece of grid paper with a very fine grid.

Consider a class of graphs that do not have a fixed planar graph as a minor. By the Grid Theorem for graphs, the members of that class have bounded branch-width. This constructive characterization provides considerable traction for both algorithmic and structural problems. For example, Robertson and Seymour [26] prove that any class of graphs of bounded branch-width is well-quasi-ordered. This extends to matroids over finite fields.

**Theorem** ([12]). *Let $\mathbb{F}$ be a finite field and $k$ an integer. Then each infinite set of $\mathbb{F}$-representable matroids with branch-width at most $k$ has two members such that one is isomorphic to a minor of the other.*

Johnson, Robertson, and Seymour [21] conjectured that also the Grid Theorem for graphs extends to matroids over finite fields and this is indeed the case.

**The Grid Theorem for matroids** ([13]). *For each finite field $\mathbb{F}$ and each planar graph $H$, there exists an integer $k$ such that each $\mathbb{F}$-representable matroid with branch-width at least $k$ has a minor isomorphic to $M(H)$.*

As a consequence we obtain the following partial result towards the Well-Quasi-Ordering Conjecture.

**Corollary.** *Let $\mathbb{F}$ be a finite field and $H$ a planar graph. Then any infinite set of $\mathbb{F}$-representable matroids with no minor isomorphic to $M(H)$ contains two matroids such that one is isomorphic to a minor of the other.*

In combination with results of Hliněný [18], we also obtain partial progress towards the Minor-Recognition Conjecture.

**Corollary.** *For each finite field $\mathbb{F}$ and each planar graph $H$, there is a polynomial-time algorithm for testing whether or not an $\mathbb{F}$-representable matroid contains a minor isomorphic to $M(H)$.*

So for matroids over a fixed finite field we can efficiently test all minor-closed properties that do not hold for the cycle matroid of all planar graphs.

Geelen and Whittle [17] show that for a finite field $\mathbb{F}$ and integer $k$, the number of excluded minors for $\mathbb{F}$-representability that have branch width at most $k$ is finite. In combination with the Grid Theorem for matroids this yields the following result.

**Corollary.** *For each finite field $\mathbb{F}$ and each planar graph $H$, there are only finitely many excluded minors for $\mathbb{F}$-representability that do not have $M(H)$ as a minor.*

We see that the structure imposed on a class of matroids by excluding the matroid of a planar graph as a minor yields restricted solutions to the Well-Quasi-Ordering Conjecture, the Minor-Recognition Conjecture, and Rota's Conjecture, and that is a promising beginning.

The Grid Theorem for matroids is absolutely central in developing a structure theory for matroids. When specialized to graphs, the proof in [13] is different from the existing proofs in [3], [27], [32]. It is important to note that we had access to an extraordinary 150-page handwritten manuscript [21] of Johnson, Robertson, and Seymour describing their progress towards a grid theorem for matroids. The techniques we learned from their manuscript played a crucial role in parts of our proof. The proof also makes use of earlier results we obtained together with Neil Robertson [10], [11].

Regarding the result above on well-quasi-ordering of $\mathbb{F}$-representable matroids of bounded branch-width it is interesting to note that the finiteness of $\mathbb{F}$ is essential there. This is illustrated by the sequence of matrices given in Section 1, below the Well-Quasi-Ordering Conjecture; they all have branch-width at most 3, as they all have rank 3. On the other hand, there are only finitely many excluded minors for the class of all matroids of branch width at most $k$, representable or not [11].

We conclude this section with a comment regarding Rota's Conjecture. We have seen that for every finite field $\mathrm{GF}(q)$ there are a finite number of excluded minors for $\mathrm{GF}(q)$-representability of any given branch width. In [15] it is proved that an excluded minor for $\mathrm{GF}(q)$-representability of sufficiently large branch width cannot contain a $\mathrm{PG}(q+6, q)$-minor. ($\mathrm{PG}(n, q)$ is the matroid represented by points of the projective geometry of order $n$ over $\mathrm{GF}(q)$.) So it follows that if Rota's Conjecture fails for $\mathrm{GF}(q)$, then there must exist excluded minors with arbitrarily large grid-minors and no large projective geometry as a minor.

## 3. An example: the structure of regular matroids

With the results in the previous section in hand we proceed towards a structure theory for matroids over finite fields. One of the prototypes of structural matroid theory and its algorithmic consequences concerns the totally unimodular matrices mentioned in Section 1. The question if a certain given matrix is totally unimodular can be translated into the question if a related, easy-to-construct, GF(2)-representable matroid is representable over all fields. Such matroids are called *regular*. Regularity is a minor-closed property. Tutte [40] proved that a GF(2)-representable matroid is regular if and only if it does not have a minor isomorphic to PG(2, 2), also called the *Fano matroid*, or to the dual of PG(2, 2). Here the *dual* of a matroid $M$ is the matroid $M^*$ with the same ground set $E$ as $M$ and with rank-function $r_{M^*}(X) = |X| - r_M(E-X) + r_M(E)$. Representability over a field $\mathbb{F}$ is closed under duality, hence so is regularity. Taking minors commutes with duality; although the roles of deletion and contraction swap.

Tutte's excluded minor characterization of regular matroids is one of the gems of matroid theory, but it does not tell how to decide if a given GF(2)-representable matroid is regular or not. That question was answered by a structural result, Seymour's Regular Matroid Decomposition Theorem [36]: *A matroid over* GF(2) *is regular if and only if it is the* 1*-,* 2*- or* 3*-sum of graphic matroids, duals of graphic matroids and copies of a particular* 10*-element matroid called* $R_{10}$. Here a 1-, 2- or 3-sum of two (representable) matroids is carried out by embedding each of them in a distinct projective space and then combining these projective spaces by taking either their direct sum, in case of a 1-sum, or identifying single points or lines, in case of a 2-sum or a 3-sum. These "meeting" points or lines should be in both matroids and may or may not be deleted from the matroid after the composition. The sums as well as the reversed "decomposition" operations preserve regularity.

So Seymour's result gives a structural description of the class of regular matroids. Their "global structure" is that they are composed from smaller pieces along low-order separations. The pieces sit together in a tree-like fashion. The description of these pieces provides the "local structure": each piece is either a graphic matroid, the dual of a graphic matroid or isomorphic to $R_{10}$. This combination of global and local structure is typical for all structural results in this paper.

Seymour's structural characterization of regular matroids is constructive, it can be used to design an algorithm for testing regularity in polynomial time. This goes as follows. First decide if the matroid is a 1-, 2- or 3-sum of smaller matroids. This can be done in polynomial time, as gluing two matrices together leaves a separation of order at most 3 in the composed matroid and Cunningham and Edmonds [2] observed that detecting these is a matroid intersection problem, which is solvable in polynomial time (Edmonds [7]). When the matroid is fully decomposed into "4-connected pieces", each piece is tested for being isomorphic to $R_{10}$, which is trivial, or being a graphic matroid or the dual of a graphic matroid, which can be done by Tutte's polynomial-time algorithm for testing graphicness [41]. If all pieces pass the test, the original matroid is regular, otherwise it is not.

By the relation between regularity and total unimodularity this yields an algorithm for testing if a real matrix is totally unimodular or not (see Schrijver [35, Chapter 20] for a description of this algorithm in terms of the matrices). This is the only known polynomial-time algorithm for testing totally unimodularity. Thus the structure of matroids is crucial for the algorithmic aspects of this central property in operations research and combinatorial optimization. Actually matroids in general do play a major role in the theory of combinatorial optimization, see Schrijver [34]. A book on matroid decomposition is Truemper [39], it mainly concerns regular matroid decomposition and related topics.

With Seymour's regular matroid decomposition in mind we next discuss what we expect to be the structure of minor-closed classes of matroids that are representable over a finite field. It should be noted that the results will not be as "tight" as in Seymour's decomposition theorem. Seymour provided a constructive description of all binary matroids that contain neither the Fano matroid nor its dual as a minor. More-

over, none of the matroids obtained via that construction contain the Fano matroid or its dual. In contrast, the Grid Theorem for graphs provides a construction for the graphs that do not contain a given planar graph $H$ as a minor, but some graphs obtained via the construction may contain $H$ as a minor. The construction is, however, sufficiently restrictive that it does not build all planar graphs. For algorithmic and well-quasi-ordering purposes, this is good enough.

## 4. Global structure and local structure

In Section 2 we discussed the structure of classes of matroids over finite fields that do not have a minor isomorphic to the cycle matroid of a particular planar graph: they can be decomposed into small pieces along low-order separations; they have low branch-width. So explorations beyond that concern matroids with high branch-width. The existence of large grid-minors in such matroids is useful in investigating their structure, but a matroid may have several high branch-width parts that are separated by low-order separations and we have to describe the structure of these parts separately. To get a handle on these parts, Robertson and Seymour [28] introduce tangles. A tangle really just indicates for each low-order separation on which side a particular high branch-width part lies. Formally, a *tangle of order* $t$ assigns to each separation $(X, Y)$ of order less than $t$ one of $X$ and $Y$ as the *small side of* $(X, Y)$ and the other side as the *big side of* $(X, Y)$. It is required that no three small sides of the tangle cover the ground set of the matroid and that no singletons are big. It turns out that the maximum order of a tangle in a matroid is the branch-width of the matroid; for graphs this was shown by Robertson and Seymour [28] and for matroids by Dhamatilike [5] (although this result was implicit in [28]).

Combining this with the Grid Theorem, we see that grid-minors yield tangles. Indeed, if $F \subseteq M$ is the set of elements of an $n$ by $n$ grid-minor of $M$, then $F$ partitions naturally in the "horizontal" and "vertical" *lines* of the grid. If we consider for each separation $(X, Y)$ of order less than $n$ the side that contains a line in $F$ as big, then that yields a tangle of order $n$. It was shown in [16] that for any finite field $\mathbb{F}$ and any $n$ there exists an integer $t$ such that in any $\mathbb{F}$-representable matroid any tangle of order $t$ controls an $n$ by $n$ grid-minor. This generalizes a result from [32] for graphs. Here a tangle *controls* a minor $N$ of $M$ if no small side of a separation of order less than the rank of $N$ contains all elements of $N$.

So tangles "locate" highly connected areas of the matroid. If all small sides of one tangle are small in some other tangle, then they both seem to refer to the same highly connected part but the latter tangle does that more accurately. Therefore we are mainly interested in the *maximal* tangles, those for which the collection of small sets is inclusion-wise maximal. It turns out that matroids, like graphs [28], can be viewed as consisting of their maximal tangles put together in a tree-shaped structure, see [16] for details. This provides a global picture of a matroid. To complete that picture we have to describe the individual tangles, the local structure.

To explain what we mean with that, we first explain what it means to reduce a set $S$ in an $\mathbb{F}$-represented matroid $M$. Consider $M$ as a collection $E$ of points in a projective geometry. Let $X$ be the span of $E - S$ in that projective geometry. For each $S' \subseteq S$ whose span meets $X$ in a single point, we call that point $x_{S'}$. Let $Y$ be the set of all points $x_{S'}$ for all such sets $S'$. Replacing $M$ by the matroid represented by the union of $Y$ and $E - S$ is called *reducing* $S$. (For graphs, this more or less means to remove the edges in $S$ and to add an edge between any pair of vertices that both lie in $S$ and in the complement of $S$.)

Let $\mathcal{C}$ be a class of matroids. A tangle has *local structure in* $\mathcal{C}$ if there exist separations $(S_1, B_1), \ldots, (S_k, B_k)$ in $M$ with disjoint small sides $S_1, \ldots, S_k$ such that the matroid obtained from $M$ by reducing each of $S_1, \ldots, S_k$ is in $\mathcal{C}$. To describe the full structure we only need to characterize the minor-closed classes $\mathcal{C}$ that provide the local structure of tangles in matroids over the finite field that do not contain a particular minor.

## 5. The local structure of graph tangles

The Graph Minor Structure Theorem says that for any $n$ there exists a surface $\Sigma$ and integers $m, d, k$ such that the tangles of a graph with no minor isomorphic to $K_n$ have local structure in the class of graphs that lie on a surface $\Sigma$ with $m$ vortices of depth at most $d$ and $k$ extra vertices. We explain what this means.

A *vortex with connectors* $v_1, \ldots, v_p$ is a graph $H$ that is the union of graphs $H_1, \ldots, H_p$ such that $v_i$ is a vertex in $H_i$ for each $i = 1, \ldots, p$ and such that if a vertex $v$ of $H$ occurs in $H_i$ and $H_j$ for some $i, j = 1, \ldots, p$ then $v$ either occurs in all of $H_{i+1}, \ldots, H_{j-1}$ or in all of $H_{j+1}, \ldots, H_{i-1}$ (indices modulo $n$). The maximum size of the subgraphs $H_1, H_2, \ldots, H_p$ is the *depth* of the vortex.

A graph is *on a surface* $\Sigma$ *with m vortices of depth at most d* if it can be constructed as follows: take a graph drawn on $\Sigma$, select $m$ faces and add to each of these faces a vortex of depth $d$ to $G$ that meets $G$ and the other added vortices only in its connectors $v_1, \ldots, v_n$ which lie in that order around the boundary of the face. If we additionally add $k$ new vertices and new edges from these vertices to each other and to the rest of the graph, we obtain *a graph that lies on a surface* $\Sigma$ *with m vortices of depth at most d and k extra vertices*.

## 6. The local structure of matroid tangles

What are the minor-closed classes needed to describe the local structure of matroids that are representable over a finite field? One natural minor-closed class is the class of graphic matroids. Also, if $\mathbb{F}'$ is a subfield of $\mathbb{F}$, then the class of $\mathbb{F}'$-representable matroids is a minor-closed class of $\mathbb{F}$-representable matroids. There is another natural class, of Dowling matroids. They are like graphs and originally introduced by Dowling [6] and studied in greater depth by Zaslavsky [44], [45].

A *Dowling matroid* is a matroid that can be represented over a field $\mathbb{F}$ by a matrix with the property that every column has at most two non-zero elements. We call such matrix a *Dowling representation* of the matroid. If the ratio between the non-zero elements in each column of a Dowling representation is in a subgroup $\Gamma$ of the multiplicative group of $\mathbb{F}$, we call the matroid *a Dowling matroid over* $\Gamma$. One can naturally associate a graph $G(A)$ with a Dowling representation $A$. Each row of $A$ is a vertex of $G(A)$ and each column of $A$ with two non-zeroes yields an edge in $G(A)$ connecting the vertices corresponding to the rows that have the non-zeroes in that column. Thus we get for each surface and each subgroup of the multiplicative group of $\mathbb{F}$ the class of $\mathbb{F}$-representable Dowling matroids that have Dowling representations over the subgroup and whose associated graphs embed on the surface. Obviously such a class is minor-closed.

In fact, we can extend such minor-closed class by allowing a bounded number of "vortices" of bounded depth, these are obtained by adding matroid elements into bounded-rank subspaces arranged in a cyclic manner around a face in the embedding, similar to vortices in graphs.

Finally we can extend a minor-closed class $\mathcal{C}$ of matroids by considering for some integer $k$ the class of all rank-$l$ perturbations of the members of $\mathcal{C}$ with $l \leq k$. Here an $\mathbb{F}$-representable matroid $M$ is a *rank-$l$ perturbation* of an $\mathbb{F}$-representable matroid $N$ if $M$ and $N$ have representations $A$ and $B$, respectively, with the linear rank of $A - B$ equal to $l$.

Splitting a vertex in a graph amounts to a rank-1 perturbation of its cycle matroid. So adding $k$ vertices to a graph amounts to adding a single vertex followed by a rank-$(k-1)$ perturbation of the cycle matroid. Adding a single vertex to a graph $G$ does in general not correspond to a low-rank perturbation. However, fortunately, the cycle matroid of the resulting graph has a Dowling representation $A$ with $G(A) = G$. Hence the Graph Minors Structure Theorem is captured by the matroid classes given above.

Now we state our main results and conjectures on the structure of minor-closed classes over a finite field GF($q$), where $q = p^k$ for some fixed prime $p$ and some fixed integer $k$. We distinguish between three types of minor-classes. The first type are the classes that do not contain the cycle matroid of large complete graphs nor their duals. The second type are the classes that do not contain large projective geometries over the prime field GF($p$) of GF($q$). The third type are the classes that do not contain large projective geometries over GF($q$). In each of the cases, $\mathcal{T}$ is a tangle in a GF($q$)-representable matroid.

Below $n$ is a fixed integer and each of the qualitative bounds "low", "bounded", or "sufficiently large" indicates a bound only depending on $q$ and $n$, so not on the particular tangles or matroids.

**Excluding $M(K_n)$ and $M(K_n)^*$.** We believe that we have proved that if $\mathcal{T}$ has sufficiently large order and does not control a minor isomorphic to $M(K_n)$ or to $M(K_n)^*$, then $\mathcal{T}$ has local structure in the class of low-rank perturbations of GF($q$)-

representable matroids that can be obtained by adding a bounded number of vortices of bounded depth to a Dowling matroid whose associated graph is embedded in a surface of low genus, or of the duals of such matroids. This implies the Graph Minors Structure Theorem.

With this result and duality, we can now restrict our attention to tangles that control the cycle matroid of a large complete graph.

**Excluding PG($n$, $p$).** We conjecture that if $\mathcal{T}$ controls a minor isomorphic to $M(K_m)$ for a sufficiently large integer $m$ but $\mathcal{T}$ does not control a minor isomorphic to PG($n$, $p$), then $\mathcal{T}$ has local structure in the class of low-rank perturbations of GF($q$)-representable Dowling matroids.

Roughly speaking the conjectures above state that if $M$ is a GF($q$)-representable matroid with no minor isomorphic to PG($n$, $p$), then $M$ admits a tree-like decomposition such that each part is either essentially a Dowling matroid or is essentially the dual of a Dowling matroid. For a field of prime order this would give the required constructive structural characterization of the minor-closed proper subclasses of matroids representable over the field.

It is interesting to note here that a slight extension of Seymour's regular matroid decomposition says that if a GF(2)-representable matroid has no minor isomorphic to PG(2, 2) then it can be constructed from graphic matroids, their duals, and copies of $R_{10}$ and copies of the dual of PG(2, 2), by 1-, 2- and 3-sums [36]. As graphic matroids are GF(2)-representable Dowling matroids and as $R_{10}$ and the dual of PG(2, 2) are low-rank perturbations of a trivial matroid, this result of Seymour's implies the conjecture above for case that $q = 2$ and $n = 2$.

**Excluding PG($n$, $q$).** For the case that $q$ is not prime, we conjecture that if $\mathcal{T}$ controls a minor isomorphic to PG($m$, $p$) for a sufficiently large integer $m$ but $\mathcal{T}$ does not control a minor isomorphic to PG($n$, $q$), then $\mathcal{T}$ has local structure in the class of GF($q$)-representable low-rank perturbations of matroids that are representable over a proper subfield of GF($q$).

Finally we can summarize all of the above into a single conjecture. For any minor-closed proper subclass $\mathcal{M}$ of GF($q$)-representable matroids, each matroid in $\mathcal{M}$ admits a tree-like decomposition such that each part is either essentially a Dowling matroid, or is essentially the dual of a Dowling matroid, or is essentially represented over a proper subfield of GF($q$).

## References

[1] Bixby, R. E., On Reid's characterization of the ternary matroids. *J. Combin. Theory Ser. B* **26** (1979), 174–204.

[2] Cunningham, W. H., and Edmonds, J., A combinatorial decomposition theory. *Canad. J. Math.* **32** (1980), 734–765.

[3] Diestel, Reinhard, Gorbunov, Konstantin Yu., Jensen, Tommy R., and Thomassen, Carsten, Highly connected sets and the excluded grid theorem. *J. Combin. Theory Ser. B* **99** (1999), 61–73.

[4] Dirac, G. A., A property in 4-chromatic graphs and some remarks on critical graphs. *J. London Math. Soc.* **27** (1952), 85–92.

[5] Dharmatilake, J., A min-max theorem using matroid separations. In *Matroid Theory* (Seattle WA 1995), Contemp. Math. 197, Amer. Math. Soc., Providence, RI, 1996, 333–342.

[6] Dowling, T., A class of geometric lattices based on finite groups. *J. Combin. Theory Ser. B* **14** (1973), 61–86.

[7] Edmonds, J., Submodular functions, matroids, and certain polyhedra. In *Combinatorial Structures and Their Applications* (Proceedings Calgary International Conference on Combinatorial Structures and Their Applications, Calgary, Alberta, 1969), ed. by R. Guy, H. Hanani, N. Sauer, and J. Schőnheim, Gordon and Breach, New York 1970, 69–87.

[8] Garey, M. R., and Johnson, D. S., *Computers and intractability: A guide to the theory of NP-completeness*, Freeman, San Fransisco, 1979.

[9] Geelen, J. F., Gerards, A. M. H., and Kapoor, A., The excluded minors for GF(4)-representable matroids. *J. Combin. Theory Ser. B* **79** (2000), 247–299.

[10] Geelen, J. F., Gerards, A. M. H., Robertson, N., and Whittle, G. P., Obstructions to branch-decomposition of matroids. Research Report 03-2, School of Mathematical and Computing Sciences, Victoria University of Wellington, 2004.

[11] Geelen, J. F., Gerards, A. M. H., Robertson, N., and Whittle, G. P., On the excluded minors for the matroids of branch-width $k$. *J. Combin. Theory Ser. B* **88** (2003), 261–265.

[12] Geelen, James F., Gerards, A. M. H., and Whittle, Geoff, Branch width and well-quasi-ordering in matroids and graphs. *J. Combin. Theory Ser. B* **84** (2002), 270–290.

[13] Geelen, Jim, Gerards, Bert, and Whittle, Geoff, Excluding a planar graph from GF($q$)-representable matroids. Research Report 03-4, School of Mathematical and Computing Sciences, Victoria University of Wellington, 2003.

[14] Geelen, Jim, Gerards, Bert, and Whittle, Geoff, Inequivalent representations of matroids I: An overview. In preparation, 2005.

[15] Geelen, Jim, Gerards, Bert, and Whittle, Geoff, Inequivalent representations of matroids II: $k$-coherent matroids. In preparation, 2004.

[16] Geelen, Jim, Gerards, Bert, and Whittle, Geoff, Tangles, tree-decompositions, and grids in matroids. Research Report 04-5, School of Mathematical and Computing Sciences, Victoria University of Wellington, 2004.

[17] Geelen, J., and Whittle, G., Branch width and Rota's conjecture. *J. Combin. Theory Ser. B* **86** (2002), 215–330.

[18] Hliněný, P., Branch-width, parse trees, and monadic second-order logic for matroids. Preprint, 2002.

[19] Hliněný, P., and Whittle, G., Matroid tree width. *European J. Combin.*, to appear.

[20] Hoffman, A. J., and Kruskal, J. B., Integral boundary points of convex polyhedra. In *Linear Inequalities and Related Systems* (ed. by H. W. Kuhn and A. W. Tucker), Ann. of Math. Stud. 38, Princeton University Press, Princeton, N.J., 1956, 223–246.

[21] Johnson, Thor, Roberston, Neil, and Seymour, P. D., Connectivity in binary matroids. Manuscript.

[22] Kuratowski, C., Sur le problème des courbes gauches en topologie. *Fund. Math.* **15** (1930), 271–283.

[23] Lazarson, T., The representation problem for independence functions. *J. London Math. Soc.* **33** (1958), 21–25.

[24] Menger, K., Zur allgemeinen Kurventheorie. *Fund. Math.* **10** (1927), 96–115.

[25] Oxley, J. G., *Matroid theory*. Oxford University Press, New York, 1992.

[26] Robertson, Neil, and Seymour, P. D., Graph Minors. IV. Tree-width and well-quasi-ordering. *J. Combin. Theory Ser. B* **48** (1990), 227–254.

[27] Robertson, Neil, and Seymour, P. D., Graph Minors. V. Excluding a planar graph. *J. Combin. Theory Ser. B* **41** (1986), 92–114.

[28] Robertson, Neil, and Seymour, P. D., Graph Minors. X. Obstructions to tree-decomposition. *J. Combin. Theory Ser. B* **52** (1991), 153–190.

[29] Robertson, Neil, and Seymour, P. D., Graph Minors. XIII. The disjoint paths problem. *J. Combin. Theory Ser. B* **63** (1995), 65–110.

[30] Robertson, Neil, and Seymour, P. D., Graph Minors. XVI. Excluding a non-planar graph. *J. Combin. Theory Ser. B* **89** (2003), 43–76.

[31] Robertson, Neil, and Seymour, P. D., Graph Minors. XX. Wagner's conjecture. *J. Combin. Theory Ser. B* **92** (2004), 325–357.

[32] Robertson, Neil, Seymour, Paul, and Thomas, Robin, Quickly excluding a planar graph. *J. Combin. Theory Ser. B* **62** (1994), 323–348.

[33] Rota, G.-C., Combinatorial theory, old and new. In *Actes du Congrès International de Mathématiciens* (Nice, 1970), vol. 3, Gauthier-Villars, Paris 1970, 229–233.

[34] Schrijver, A., *Combinatorial optimization — polyhedra and efficiency*. Vol. A, Algorithms Combin. 24, Springer-Verlag, Berlin 2003.

[35] Schrijver, A., *Theory of linear and integer programming*. John Wiley and Sons, Chichester 1986.

[36] Seymour, P. D., Decomposition of regular matroids. *J. Combin. Theory Ser. B* **28** (1980), 305–359.

[37] Seymour, P. D., Matroid representation over GF(3). *J. Combin. Theory Ser. B* **26** (1979), 159–173.

[38] Seymour, P. D., Recognizing graphic matroids. *Combinatorica* **1** (1981), 75–78.

[39] Truemper, K., *Matroid decomposition*. Academic Press, San Diego, 1992.

[40] Tutte, W. T., A homotopy theorem for matroids I, II. *Trans. Amer. Math. Soc.* **88** (1958), 144–174.

[41] Tutte, W. T., An algorithm for determining whether a given binary matroid is graphic. *Proc. Amer. Math. Soc.* **11** (1960), 905–917.

[42] Tutte, W. T., Menger's theorem for matroids. *J. Res. Nat. Bur. Standards Sect. B.* **69B** (1965), 49–53.

[43] Welsh, D. J. A., *Matroid theory*, Academic Press, London, 1976.

[44] T. Zaslavsky, A mathematical bibliography of signed and gain graphs and allied areas. Manuscript prepared with Marge Pratt, *Electron. J. Combin.* **5** (1999), Dynamic Surveys 8, 124 pp.

[45] Zaslavsky, T., Biased graphs. II. the three matroids. *J. Combin. Theory Ser. B* **51** (1991), 46–72.

Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada
E-mail: jfgeelen@uwaterloo.ca

Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands
and
Technische Universiteit Eindhoven, Eindhoven, The Netherlands
E-mail: bert.gerards@cwi.nl

School of Mathematical and Computing Sciences, Victoria University, Wellington, New Zealand
E-mail: geoff.whittle@vuw.ac.nz

# Cherednik algebras, Macdonald polynomials and combinatorics

Mark Haiman*

**Abstract.** In the first part of this article we review the general theory of Cherednik algebras and non-symmetric Macdonald polynomials, including a formulation and proof of the fundamental *duality theorem* in its proper general context. In the last section we summarize some of the combinatorial results in this area which we have recently obtained in collaboration with J. Haglund and N. Loehr.

**Mathematics Subject Classification (2000).** Primary 33D52; Secondary 05E10.

**Keywords.** Macdonald polynomials, affine Hecke algebras, Cherednik algebras.

## 1. Introduction

*The record is very long. The facts are few and may be briefly stated.*
*—Miller v. San Francisco Methodist Episcopal* (1932)

This article consists of an overview of the theory of Cherednik algebras and non-symmetric Macdonald polynomials, followed by the combinatorial formula for non-symmetric Macdonald polynomials of type $A_{n-1}$ recently obtained by Haglund, Loehr and the author.

The main points in the theory are duality (Theorems 4.10, 5.11), and its consequence, the intertwiner recurrence for Macdonald polynomials (Corollary 6.15), which is the key to the combinatorial study of non-symmetric Macdonald polynomials. The intertwiner recurrence can also be used to deduce other important results in the theory, such as the norm and evaluation formulas, but I have omitted those for lack of space.

The theory of course did not spring into being in the tidy form in which I have attempted to package it here. Rather, it has been gradually clarified over almost twenty years through the efforts of many people, in a large literature which I will not attempt to cite in full. Let me only mention the origins of the theory in the works of Macdonald [13], [14], [15], Opdam [17], and Cherednik [1], [2] and remark that further important contributions were made by Ion, Knop, Koornwinder, Sahi, and van Diejen, among others.

*Work supported in part by NSF grant DMS-0301072.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The overview given here necessarily has much in common with Macdonald's monograph [16], which serves a similar purpose, but there are also several differences. I have systematically used the lattice formulation for root systems, because it is most natural from related points of view (algebraic groups, quantum groups), because it puts affine and other root systems on an equal footing, and because important elements of the theory (§2, 5.1–5.5, 5.13–5.15) apply to arbitrary root systems. I give a new and somewhat more general proof of the duality theorem; Macdonald's proof, strictly speaking, applies to the root system of $\mathrm{SL}_n$, for instance, but not $\mathrm{GL}_n$ or $\mathrm{PGL}_n$, although it can be adjusted to cover these cases. For the triangularity property of the Macdonald polynomials $E_\lambda$ (Theorem 6.6), I use the affine Bruhat order on the weight lattice $X$, rather than the orbit-lexicographic order used by Macdonald. This simplifies some arguments, and is more natural in that the coefficient of $x^\mu$ in $E_\lambda$ is non-zero if and only if $\mu < \lambda$ in Bruhat order. I have also tried to use more transparent notation.

## 2. Root systems

**2.1.** We always consider root systems realized in a lattice. So, for us, a *root system* $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ consists of a finite-rank free abelian group $X$, whose dual lattice $\mathrm{Hom}(X, \mathbb{Z})$ is denoted $X^\vee$, a finite set of vectors $\alpha_1, \dots, \alpha_n \in X$, called *simple roots*, and a finite set of covectors $\alpha_1^\vee, \dots, \alpha_n^\vee \in X^\vee$, called *simple coroots*. We denote by $X_\mathbb{Q}$ (resp. $X_\mathbb{R}$) the $\mathbb{Q}$-vector space $X \otimes_\mathbb{Z} \mathbb{Q}$ (resp. $\mathbb{R}$-vector space $X \otimes_\mathbb{Z} \mathbb{R}$) spanned by $X$.

The $n \times n$ matrix $A$ with entries $a_{ij} = \langle \alpha_j, \alpha_i^\vee \rangle$ is assumed to be a *generalized Cartan matrix*, satisfying the axioms

$$\begin{aligned} &\text{(i)} && \langle \alpha_i, \alpha_i^\vee \rangle = 2, \\ &\text{(ii)} && \langle \alpha_j, \alpha_i^\vee \rangle \le 0 \text{ for all } j \ne i, \\ &\text{(iii)} && \langle \alpha_j, \alpha_i^\vee \rangle = 0 \text{ if and only if } \langle \alpha_i, \alpha_j^\vee \rangle = 0. \end{aligned}$$

The *Dynkin diagram* is the graph with nodes $i = 1, \dots, n$ and an edge $\{i, j\}$ for each $a_{ij} \ne 0$, usually with some decoration on the edges to indicate the values of $a_{ij}, a_{ji}$. If the Dynkin diagram is connected, $A$ is *indecomposable*. If there exist non-zero integers $d_i$ such that $\langle \alpha_j, d_i \alpha_i^\vee \rangle = \langle \alpha_i, d_j \alpha_j^\vee \rangle$ for all $i$, $j$, then $A$ is *symmetrizable*. The integers $d_i$ can be assumed positive. If $A$ is symmetrizable and indecomposable, the $d_i$ are unique up to an overall common factor. Then $d_i$ is *length* of the root $\alpha_i$. If there are only two root lengths, we call them *long* and *short*. If there is only one root length, every root is both long and short.

**2.2.** Let $\alpha \in X$ and $\alpha^\vee \in X^\vee$ satisfy $\langle \alpha, \alpha^\vee \rangle = 2$. The linear automorphism

$$s_{\alpha\alpha^\vee}(\lambda) = \lambda - \langle \lambda, \alpha^\vee \rangle \, \alpha$$

of $X$ is a *reflection*. It fixes the hyperplane $\langle \lambda, \alpha^\vee \rangle = 0$ pointwise, and sends $\alpha$ to $-\alpha$. Thus $(s_{\alpha,\alpha^\vee})^2 = 1$. The reflection on $X^\vee$ dual to $s_{\alpha,\alpha^\vee}$ is equal to $s_{\alpha^\vee,\alpha}$.

If $\alpha^\vee$ is implicitly associated to $\alpha$, we write $s_\alpha$ for both $s_{\alpha,\alpha^\vee}$ and $s_{\alpha^\vee,\alpha}$. When $\alpha = \alpha_i$ and $\alpha^\vee = \alpha_i^\vee$ are a simple root and corresponding coroot, we write $s_i$ for $s_{\alpha_i}$. The $s_i$ are called *simple reflections*.

**2.3.** The root system $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is *non-degenerate* if the simple roots $\alpha_i$ are linearly independent. When the Cartan matrix $A$ is non-singular, *e.g.*, for any finite root system, then both $X$ and its dual $\big(X^\vee, (\alpha_i^\vee), (\alpha_i)\big)$ are necessarily non-degenerate. When $A$ is singular, for instance if the root system is affine (Definition 3.1), it is often convenient to take the simple roots to be a basis of $X_{\mathbb{Q}}$, in which case $X$ is non-degenerate but its dual is degenerate.

**2.4.** Assume in what follows that $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is non-degenerate. The *Weyl group* $W$ is the group of automorphisms of $X$ (and of $X^\vee$) generated by the simple reflections $s_i$. The sets of *roots* and *coroots* are

$$R = \bigcup_i W(\alpha_i), \quad R^\vee = \bigcup_i W(\alpha_i^\vee).$$

The *root* and *coroot lattices* are

$$Q = \mathbb{Z}\{\alpha_1, \ldots, \alpha_n\} \subseteq X, \quad Q^\vee = \mathbb{Z}\{\alpha_1^\vee, \ldots, \alpha_n^\vee\} \subseteq X^\vee.$$

The set of *positive roots* is $R_+ = R \cap Q_+$, where

$$Q_+ = \mathbb{N}\{\alpha_1, \ldots, \alpha_n\}.$$

The *dominant weights* are the elements of the cone

$$X_+ = \{\lambda \in X : \langle \lambda, \alpha_i^\vee \rangle \geq 0 \text{ for all } i\}.$$

The root system $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is *finite* if $W$ is a finite group, or equivalently, $R$ is a finite set. The Cartan matrix $A$ of a finite root system is symmetrizable, with positive definite symmetrization $DA$. Conversely, if $A$ has a positive definite symmetrization, then $R$ is finite. The finite root systems classify reductive algebraic groups $G$ over any algebraically closed field $k$. Then $X$ is the character group of a maximal torus in $G$, or *weight lattice*.

**Example 2.5.** Let $X = \mathbb{Z}^n$, and identify $X^\vee$ with $X$ using the standard inner product on $\mathbb{Z}^n$ such that the unit vectors $e_i$ are orthogonal. Let $\alpha_i = \alpha_i^\vee = e_i - e_{i+1}$ for $i = 1, \ldots, n-1$. This gives the root system of the group $\mathrm{GL}_n$.

Replacing $X$ with the root lattice $Q$ and restricting the simple coroots to $Q$, we obtain the root system of the adjoint group $\mathrm{PGL}_n$ ($\mathrm{GL}_n$ modulo its center).

The constant vector $\varepsilon = e_1 + \cdots + e_n$ satisfies $\langle \varepsilon, \alpha_i^\vee \rangle = 0$ for all $i$. Let $X' = X/(\mathbb{Z}\varepsilon)$, with simple roots and coroots induced by those of $X$. This gives the root system of the simply connected group $\mathrm{SL}_n$. It is dual to the root system of $\mathrm{PGL}_n$. All three root systems have the same Cartan matrix, of type $A_{n-1}$.

**2.6.** We recall some standard facts. First, $R = R_+ \cup -R_+$, *i.e.*, every root is positive or negative (note that $R = -R$, since $s_i(\alpha_i) = -\alpha_i$ for all $i$). The Weyl group $W$, with its generating set $S$ of simple reflections $s_i$, is a Coxeter group with defining relations

$$s_i^2 = 1, \tag{1}$$

$$s_i s_j s_i \ldots = s_j s_i s_j \ldots \quad (m_{ij} \text{ factors on each side}), \tag{2}$$

where if $a_{ij}a_{ji} = 0, 1, 2$ or $3$, then $m_{ij} = 2, 3, 4$, or $6$, respectively, and if $a_{ij}a_{ji} \geq 4$, there is no relation between $s_i, s_j$.

The *length* $l(w)$ of $w \in W$ is the minimal $l$ such that $w = s_{i_1} \ldots s_{i_l}$. Such an expression is called a reduced factorization. More generally, if $w = u_1 u_2 \ldots u_r$ with $l(w) = l(u_1) + \cdots + l(u_r)$ we call $u_1 \cdot u_2 \ldots u_r$ a reduced factorization.

If $w = s_{j_1} \ldots s_{j_l}$ is a second reduced factorization, then the identity $s_{j_1} \ldots s_{j_l} = s_{i_1} \ldots s_{i_l}$ holds in the monoid with generators $s_i$ and the *braid relations* (2), that is, it does not depend on the relations $s_i^2 = 1$.

The length of $w$ is equal to the number of positive roots carried into negative roots by $w$, *i.e.*, $l(w) = |R_+ \cap w^{-1}(-R_+)|$. In particular, $\alpha_i$ is the only positive root $\alpha$ such that $s_i(\alpha) \in -R_+$. The following conditions are equivalent: (i) $l(ws_i) < l(w)$; (ii) $w(\alpha_i) \in -R_+$; (iii) some reduced factorization of $w$ ends with $s_i$. We abbreviate these conditions to $ws_i < w$, and write $s_i w < w$ when $w^{-1}s_i < w^{-1}$.

If $\alpha = w(\alpha_i) = w'(\alpha_j)$, then $w(\alpha_i^\vee) = w'(\alpha_j^\vee)$, so there is a well-defined coroot $\alpha^\vee = w(\alpha_i^\vee)$ associated to $\alpha$ and satisfying $\langle \alpha, \alpha^\vee \rangle = 2$, and accordingly a well-defined reflection $s_\alpha = s_{\alpha,\alpha^\vee} = ws_i w^{-1}$. Warning: the correspondence $\alpha \mapsto \alpha^\vee$ need not be bijective if the dual root system is degenerate.

The map $W \to \{\pm 1\}$, $w \mapsto (-1)^{l(w)}$ is a group homomorphism. In particular, $l(s_\alpha)$ is always odd, and $l(ws_\alpha) \neq l(w)$. We put $ws_\alpha < w$ if $l(ws_\alpha) < l(w)$. The *Bruhat order* is the partial order on $W$ given by the transitive closure of these relations.

**2.7.** The *braid group* $\mathcal{B}(W)$ is the group with generators $T_i$ and the braid relations (2) with $T_i$ in place of $s_i$. If $w = s_{i_1} \ldots s_{i_l}$ is a reduced factorization, we set $T_w = T_{i_1} \ldots T_{i_l}$. These elements are well-defined and satisfy

$$T_u T_v = T_{uv} \quad \text{when } uv = u \cdot v \text{ is a reduced factorization.} \tag{3}$$

There is a canonical homomorphism $\mathcal{B}(W) \to W$, $T_i \mapsto s_i$. By the symmetry of the braid relations, there is an automorphism $T_i \leftrightarrow T_i^{-1}$ of $\mathcal{B}(W)$.

**2.8.** The *affine Weyl group* of $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is the semidirect product $W \ltimes X$. In this context, we use multiplicative notation for the group $X$, denoting $\lambda \in X$ by $x^\lambda$. Explicitly, $W \ltimes X$ is generated by its subgroups $W$ and $X$ with the additional relations

$$s_i x^\lambda s_i = x^{s_i(\lambda)}. \tag{4}$$

**2.9.** The (left) *affine braid group* $\mathcal{B}(W, X)$ of $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is the group generated by $\mathcal{B}(W)$ and $X$, with the additional relations

$$T_i x^\lambda = x^\lambda T_i \quad \text{if } \langle \lambda, \alpha_i^\vee \rangle = 0 \ \ (\textit{i.e.}, \text{if } s_i(\lambda) = \lambda); \tag{5}$$

$$T_i x^\lambda T_i = x^{s_i(\lambda)} \quad \text{if } \langle \lambda, \alpha_i^\vee \rangle = 1. \tag{6}$$

These two relations may be combined into the following analog of (4):

$$T_i^a x^\lambda T_i^b = x^{s_i(\lambda)}, \quad \text{where } a, b \in \{\pm 1\} \text{ and } \langle \lambda, \alpha_i^\vee \rangle = (a+b)/2 \tag{7}$$

(the case $a = b = -1$ follows by taking inverses on both sides in (6)). The canonical homomorphism $\mathcal{B}(W) \to W$ extends to a homomorphism $\mathcal{B}(W, X) \to W \ltimes X$ which is the identity on $X$.

For clarity when dealing with double affine braid groups later on, we define separately the *right* affine braid group $\mathcal{B}(X, W)$, generated by $W$ and $X$ with additional relations

$$T_i x^\lambda = x^\lambda T_i \quad \text{if } \langle \lambda, \alpha_i^\vee \rangle = 0\,; \tag{8}$$

$$T_i^{-1} x^\lambda T_i^{-1} = x^{s_i(\lambda)} \quad \text{if } \langle \lambda, \alpha_i^\vee \rangle = 1. \tag{9}$$

There is an isomorphism $\mathcal{B}(X, W) \cong \mathcal{B}(W, X)$ which maps $T_i \mapsto T_i^{-1}$ and is the identity on $X$.

**2.10.** If $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is a non-degenerate root system, the root lattice $Q$ is free with basis $(\alpha_i)$. Identify $Q^\vee$ with a quotient of the free abelian group $\hat{Q}^\vee$ with basis $(\alpha_i^\vee)$, and set $P = \operatorname{Hom}(\hat{Q}^\vee, \mathbb{Z})$. The roots and coroots in $X$ are then given by homomorphisms $Q \to X \to P$, where the matrix of the composite $Q \to P$ is the Cartan matrix $A$. Suppose that $X \to P$ factors through a second lattice $X'$ as

$$Q \to X \underset{j}{\to} X' \to P.$$

This induces a root system $\big(X', (\alpha_i'), (\alpha_i'^\vee)\big)$ in $X'$ with the same Cartan matrix $A$ and canonically isomorphic Weyl and braid groups $W' = W$, $\mathcal{B}(W') = \mathcal{B}(W)$. There is an induced homomorphism of affine braid groups $j_{\mathcal{B}} \colon \mathcal{B}(W, X) \to \mathcal{B}(W, X')$ which restricts to $j$ on $X$ and to the canonical isomorphism on $\mathcal{B}(W)$.

**Theorem 2.11.** *The image of $j_{\mathcal{B}} \colon \mathcal{B}(W, X) \to \mathcal{B}(W, X')$ is normal in $\mathcal{B}(W, X')$, and the induced maps* $\ker(j) \to \ker(j_{\mathcal{B}})$, $\operatorname{coker}(j) \to \operatorname{coker}(j_{\mathcal{B}})$ *are isomorphisms.*

*Proof* (*outline*). First suppose that $X' = X \oplus \mathbb{Z}\nu$, where $\langle \nu, \alpha_i^\vee \rangle \in \{0, 1\}$ for all $i$. One proves that there exists an automorphism $\eta$ of $\mathcal{B}(W, X)$ which fixes $X$, such that $\eta(T_i) = T_i$ if $\langle \nu, \alpha_i^\vee \rangle = 0$, and $\eta(T_i) = T_i^{-1} x^{-\alpha_i}$ if $\langle \nu, \alpha_i^\vee \rangle = 1$. Then one checks that $\eta^{\mathbb{Z}} \ltimes \mathcal{B}(W, X) \cong \mathcal{B}(W, X')$, with $\eta \mapsto x^\nu$. Iterating this gives $\mathcal{B}(W, X \oplus P) \cong P \ltimes \mathcal{B}(W, X)$, and similarly, $\mathcal{B}(W, X' \oplus P) \cong P \ltimes \mathcal{B}(W, X')$.

Replacing $X$, $X'$ with $X \oplus P$, $X' \oplus P$, we may assume that $X \to P$ and $X' \to P$ are surjective.

Next one verifies that if $X \to X'$ is surjective, with kernel $Z$, then $\mathcal{B}(W, X') \cong \mathcal{B}(W, X)/Z$. Applying this to $0 \to Z \to X \to P \to 0$ and $0 \to Z' \to X' \to P \to 0$, we get surjections $\mathcal{B}(W, X) \to \mathcal{B}(W, P)$, $\mathcal{B}(W, X') \to \mathcal{B}(W, P)$ with kernels $Z$, $Z'$. The theorem then follows by some easy diagram chasing. □

**2.12.** Let $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ be a root system. It may happen that for one or more of the simple roots $\alpha_i$, we have $\alpha_i^\vee \in 2X^\vee$. Then we can form another (degenerate) root system by adjoining a new simple root $2\alpha_i$ and coroot $\alpha_i^\vee/2$. Note that $s_{(2\alpha_i),(\alpha_i^\vee/2)} = s_i$, so this new root system has the same Weyl group as the original one, but a larger set of roots $R' = R \cup W(2\alpha_i)$.

If a root system contains two simple roots $\alpha_i, \alpha_{i'}$ such that $s_i = s_{i'}$ and $\alpha_i \neq \pm\alpha_{i'}$, it is said to be *non-reduced*, otherwise it is *reduced*. We remark that $s_{i'} = s_i$ implies $\alpha_{i'} = d\alpha_i$, $\alpha_{i'}^\vee = (1/d)\alpha_i^\vee$, where $d \in \{\pm 1, \pm 2, \pm 1/2\}$. Hence every non-reduced root system is constructed by extensions as above from a reduced root system with the same Weyl group.

## 3. Affine root systems and affine Weyl groups

**Definition 3.1.** A root system $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is *affine* if its Cartan matrix $A$ is singular, and for every proper subset $J$ of the indices, the root system $\big(X, (\alpha_i)_{i\in J}, (\alpha_i^\vee)_{i\in J}\big)$ is finite.

**3.2.** The definition implies that the nullspace of $A$ is one-dimensional. If $X$ is non-degenerate, then $\{\lambda \in Q : \langle \lambda, \alpha_i^\vee \rangle = 0 \text{ for all } i\}$ is a sublattice of rank 1. It always has a (unique) generator $\delta \in Q_+$, called the *nullroot*.

We index the simple roots by $i = 0, 1, \ldots, n$. We always assume that $i = 0$ is an *affine node*, meaning that $\alpha_0 \in \mathbb{Q}\alpha + \mathbb{Q}\delta$ for some root $\alpha$ of the finite root system $\big(X, (\alpha_1, \ldots, \alpha_n), (\alpha_1^\vee, \ldots, \alpha_n^\vee)\big)$. This condition is equivalent to $s_1, \ldots, s_n$ generating the finite Weyl group $W_0 = W/Q_0'$, where $W$ is the Weyl group and $Q_0'$ is the kernel of its induced action on $X/(X \cap \mathbb{Q}\delta)$. Every affine root system has at least one affine node.

**3.3.** The affine Cartan matrices are classified in Kac [8] and Macdonald [16]. They are symmetrizable and indecomposable. We refer to them using Macdonald's nomenclature, but with a tilde over the names to distinguish them from finite types. Those denoted $\widetilde{X}_n$, or $X_n^{(1)}$ in Kac, are the *untwisted types*, where $X_n = A_n$, $B_n$, $C_n$, $D_n$, $E_{6,7,8}$, $F_4$, or $G_2$ is a Cartan matrix of finite type. Their duals (if different) $\widetilde{B}_n^\vee$, $\widetilde{C}_n^\vee$, $\widetilde{F}_4^\vee$, $\widetilde{G}_2^\vee$ are the *dual untwisted types*, denoted $A_{2n-1}^{(2)}$, $D_{n+1}^{(2)}$, $E_6^{(2)}$, and $D_4^{(3)}$ in Kac.

The remaining *mixed* types, denoted $A_{2n}^{(2)}$ in Kac, are exceptional in that they have three root lengths. Although the mixed types are isomorphic to their duals, we prefer

to distinguish between them, denoting a mixed type as $\widetilde{BC}_n$ when the distinguished affine root $\alpha_0$ is the longest simple root, and $\widetilde{BC}_n^\vee$ when $\alpha_0$ is the shortest simple root.

Types $\widetilde{B}_n$, $\widetilde{C}_n^\vee$, $\widetilde{BC}_n$, $\widetilde{BC}_n^\vee$ contain one or more simple roots $\alpha_i$ such that $\langle \alpha_j, \alpha_i^\vee \rangle$ is even for all $j$. There exist affine root systems $X$ of these types such that $\alpha_i^\vee \in 2X^\vee$. A *non-reduced affine root system* is a non-reduced extension (§2.12) of such a root system $X$.

**3.4.** The Weyl group $W_a$ of any affine root system $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ is isomorphic to the affine Weyl group $W = Q_0' \rtimes W_0$ of some finite root system $\big(Y, (\alpha_i'), (\alpha_i'^\vee)\big)$. Conversely, the affine Weyl group $Y \rtimes W_0$ of any finite root system is a semidirect extension $\Pi \ltimes W_a$ of the Weyl group of a corresponding affine root system. We now fix precise notation and explain how this correspondence comes about.

**3.5.** Let $\big(Y, (\alpha_i'), (\alpha_i'^\vee)\big)$, $i = 1, \ldots, n$, be a finite root system, with Weyl group $W_0$ and root lattice $Q_0'$. Let $\phi'$ be the (unique) dominant short root. Let $W_e = Y \rtimes W_0$ be the affine Weyl group of $Y$, and set $W_a = Q_0' \rtimes W_0 \subseteq W_e$. Write $y^\lambda$ for $\lambda \in Y$ regarded as an element of $W_e$. The orbit $W_0(\phi')$ consists of all the short roots, and spans $Q_0$. Defining $s_0 = y^{\phi'} s_{\phi'}$, it follows that $s_0$ and $s_1, \ldots, s_n \in W_0$ generate $W_a$. We will construct an affine root system whose Weyl group $W$ is isomorphic to $W_a$, with simple reflections corresponding to the generators $s_0, \ldots, s_n$.

**3.6.** Let $X = Y^\vee \oplus \mathbb{Z}$, and fix a non-zero element $\delta$ in the second summand. We need not assume that $\delta$ is a generator, so in general we have $X = Y^\vee \oplus \mathbb{Z}\delta/m$ for some positive integer $m$. Define the pairing $\langle X, Y \rangle \to \mathbb{Z}$, extending the canonical pairing $\langle Y^\vee, Y \rangle \to \mathbb{Z}$, with $\langle \delta, Y \rangle = 0$.

Let $\theta = \phi'^\vee$ be the highest coroot. For $i \neq 0$, set $\alpha_i = \alpha_i'^\vee$ and $\alpha_i^\vee = \alpha_i'$ (regarded as a linear functional on $X$ via $\langle \cdot, \cdot \rangle$). Put $\alpha_0 = \delta - \theta$ and $\alpha_0^\vee = -\phi'$. The subgroup $W_0 \subseteq W_a$ acts via its original action on $Y^\vee$, fixing $\delta$. The subgroup $Q_0' \subseteq W_a$ acts by *translations*, given by the formula

$$y^{\beta'}(\mu^\vee) = \mu^\vee - \langle \mu^\vee, \beta' \rangle \delta, \tag{10}$$

One checks that the element $y^{\phi'} s_{\phi'} \in W_a$ acts as the simple reflection $s_0$, identifying $W_a$ with the Weyl group $W$ of $X$.

For $Y$ of type $Z_n$ ($Z = A, B, \ldots, G$), the affine root system $X$ just constructed is of untwisted type $\widetilde{Z}_n$, with nullroot $\delta$. In this case the affine roots are

$$R = R_0'^\vee + \mathbb{Z}\delta, \tag{11}$$

and the positive roots are $R_+ = (R_0'^\vee + \mathbb{Z}_{>0}\delta) \cup (R_0'^\vee)_+$.

**3.7.** Let $\big(X, (\alpha_0, \ldots, \alpha_n), (\alpha_0^\vee, \ldots, \alpha_n^\vee)\big)$ be any affine root system, $W$ its Weyl group. Let $Q_0$, $W_0$ be the root lattice and Weyl group of the finite root system $\big(X, (\alpha_1, \ldots, \alpha_n), (\alpha_1^\vee, \ldots, \alpha_n^\vee)\big)$. If $X$ is of untwisted type, we have just seen that $W \cong Q_0' \rtimes W_0$, where $Q_0' = Q_0^\vee$. If $X^\vee$ is of untwisted type, then $W \cong W(X^\vee) \cong$

$Q_0' \rtimes W_0$, where $Q_0' = (Q_0^\vee)^\vee = Q_0$. If $X$ is of mixed type, its Weyl group is of type $\widetilde{C}_n$, so $W \cong Q_0' \rtimes W_0$ where $Q_0'$ is of type $C_n$, hence $Q_0' = Q_0^\vee$ for $\widetilde{BC}_n$, and $Q_0' = Q_0$ for $\widetilde{BC}_n^\vee$.

**3.8.** Twisted affine root systems can also be constructed in the manner of §3.6, by taking $\theta$ to be any dominant coroot of $Y$ or of a non-reduced finite root system containing $Y$. This yields dual untwisted types when $\theta$ is short, and mixed types when $\theta$ is one-half of a long coroot or twice a short coroot. However, when $\theta \neq \phi'^\vee$, we no longer have $W = Q_0' \rtimes W_0$.

**3.9.** We now return to the situation of §3.5, fixing the finite root system $Y$ and untwisted affine root system $X = Y^\vee \oplus \mathbb{Z}\delta/m$ in what follows. The affine Weyl group $W_e = Y \rtimes W_0$ of $Y$ is called the *extended affine Weyl group*. The action of $Q_0'$ on $X$ given by (10) extends to an action of $Y$, hence the action of $W_a = Q_0' \rtimes W_0$ extends to $W_e$. By (11), $W_e$ preserves the set of affine roots $R$.

**3.10.** The further properties of $W_a$ and $W_e$ are best understood in terms of the following "alcove picture." Let $H = \{x \in X_{\mathbb{R}}^\vee : \langle \delta, x\rangle = 1\}$ be the *level 1 plane*, and let $\Lambda_0^\vee \in H$ be the linear functional $\Lambda_0^\vee(Y^\vee) = 0$, $\langle \delta, \Lambda_0^\vee\rangle = 1$. The group $W_e$ fixes $\delta$, hence acts on $H$. The translations $Y \subset W_e$ act on $H$ by

$$y^\lambda(\mu) = \mu + \lambda, \tag{12}$$

and the finite Weyl group $W_0$ is generated by reflections fixing $\Lambda_0^\vee$. In particular, the map $y^\lambda \mapsto \Lambda_0^\vee + \lambda$ identifies $Y \cong W_e/W_0$ with the orbit $W_e(\Lambda_0^\vee) \subset H$, equivariantly with respect to the original action of $W_0$ on $Y$, and the action of $Q_0' \subseteq Y$ by translations.

Each affine root $\alpha \in R$ induces an affine-linear functional $\alpha(x) = \langle \alpha, x\rangle$ on $H$. Its zero set $h_\alpha = \{x \in H : \alpha(x) = 0\}$ is an affine hyperplane in $H$, and $s_\alpha \in W = W_a$ fixes $h_\alpha$ pointwise. The space $H$ is tessellated by *affine alcoves* bounded by the root hyperplanes $h_\alpha$. We distinguish the *dominant alcove* $A_0 = H \cap (\mathbb{R}_+ X_+^\vee) = \{x \in H : \alpha(x) \geq 0 \text{ for all } \alpha \in R_+\}$.

The alcove $A_0$ is a fundamental domain for the action of $W_a$ on $H$. Its walls are the root hyperplanes $h_{\alpha_i}$ for the simple affine roots $\alpha_0, \ldots, \alpha_n$. Let $\Pi \subseteq W_e$ be the stabilizer of $A_0$, or equivalently, $\Pi = \{\pi \in W_e : \pi(R_+) = R_+\}$. Since $\Pi$ preserves the set of simple roots, it normalizes the subgroup $W_a \subseteq W_e$ and the set of Coxeter generators $S = \{s_0, \ldots, s_n\} \subseteq W_a$. The following are immediate.

**Corollary 3.11.** *With the notation above, we have $W_e = \Pi \ltimes W_a$. Moreover, $\Pi$ is the normalizer in $W_e$ of the set of Coxeter generators $S = \{s_0, \ldots, s_n\}$.*

**Corollary 3.12.** *The canonical homomorphism $Y \subset W_e \to W_e/W_a = \Pi$ induces an isomorphism $Y/Q_0' \cong \Pi$. In particular, $\Pi$ is abelian.*

To make this explicit, write $\pi \in \Pi$ uniquely as

$$\pi = y^{\lambda_\pi} \cdot v_\pi \in Y \rtimes W_0. \tag{13}$$

Then $\pi$ maps to the coset of $\lambda_\pi$ in $Y/Q_0'$. In the notation of §3.10, we have $\Lambda_0^\vee + \lambda_\pi = y^{\lambda_\pi}(\Lambda_0^\vee) = \pi(\Lambda_0^\vee) \in A_0$. Equivalently, $\lambda_\pi \in Y$ is a dominant weight such that $\langle \lambda, \phi'^\vee \rangle \leq 1$, or *minuscule weight*. Conversely, if $\lambda \in Y$ is minuscule, there is a unique $\pi \in \Pi$ such that $y^{\lambda - \lambda_\pi} \in W_a$. Then $\lambda = \lambda_\pi$, because both weights are minuscule and $A_0$ is a fundamental domain for $W_a$. The minuscule weights $\lambda_\pi$ (including $\lambda_1 = 0$) are thereby in bijection with $\Pi$.

**3.13.** The distinguished elements

$$y^{\phi'} = s_0 s_{\phi'}, \quad y^{\lambda_\pi} = \pi v_\pi^{-1}, \tag{14}$$

where $\phi'$ is the dominant short root and $\lambda_\pi$ are the minuscule weights, are characterized as the unique translations such that $s_0 \in y^{\phi'} W_0$, $\pi \in y^{\lambda_\pi} W_0$, consistent with our having written $W_e = Y \rtimes W_0$. If we write $W_e = W_0 \ltimes Y$, we instead distinguish the translations

$$y^{-\phi'} = s_{\phi'} s_0, \quad y^{-\lambda_\pi} = v_\pi \pi^{-1} \tag{15}$$

corresponding to the *anti-dominant* short root and the "anti-minuscule" weights. Of course (14) and (15) are equivalent, but the corresponding formulas for the left and right affine braid groups will not be (see Theorem 4.2, Corollary 4.3).

## 4. Double affine braid groups

**4.1.** Let $W_e = Y \rtimes W_0 = \Pi \ltimes W_a$ be an extended affine Weyl group (§3.9–3.13). By Corollary 3.11, $\Pi$ acts on $W_a$ by Coxeter group automorphisms. Hence $\Pi$ also acts on $\mathcal{B}(W_a)$, and we can form the *extended affine braid group* $\mathcal{B}(W_e) = \Pi \ltimes \mathcal{B}(W_a)$.

Define the length function on $W_e = \Pi \ltimes W_a$ by $l(\pi w) = l(w)$. Note that $l(w\pi) = l(\pi w^\pi) = l(w^\pi) = l(w)$. The length of $v = \pi w$ is again equal to $|R_+ \cap v^{-1}(-R_+)|$, or to the number of affine hyperplanes $h_\alpha$ separating $v(A_0)$ from $A_0$ in the alcove picture (§3.10). Identity (3) continues to hold in $\mathcal{B}(W_e)$.

The counterpart to Corollary 3.11 is the following theorem of Bernstein (see [9, (4.4)]).

**Theorem 4.2.** *The identification $\Pi \ltimes W_a = Y \rtimes W_0$ lifts to an isomorphism $\mathcal{B}(W_e) \cong \mathcal{B}(Y, W_0)$ between the extended affine braid group defined above, and the* (*right*) *affine braid group* (§2.9) *of the finite root system $Y$. The isomorphism is the identity on $\mathcal{B}(W_0)$ and given on the remaining generators by $y^{\phi'} \leftrightarrow T_0 T_{s_{\phi'}}$, $y^{\lambda_\pi} \leftrightarrow \pi T_{v_\pi^{-1}}$, in the notation of* §3.5 *and* (13).

We describe the restriction of the isomorphism to $Y \subseteq \mathcal{B}(Y, W_0)$ more explicitly. If $\lambda, \mu \in Y_+$ are dominant, the alcove picture shows that $l(y^{\lambda+\mu}) = l(y^\lambda) + l(y^\mu)$. Hence $T_{y^{\lambda+\mu}} = T_{y^\lambda} T_{y^\mu}$ in $\mathcal{B}(W_e)$. It follows that there is a well-defined group homomorphism $\phi\colon Y \to \mathcal{B}(W_e)$ such that $y^{\lambda-\mu} \mapsto T_{y^\lambda} T_{y^\mu}^{-1}$ for $\lambda, \mu \in Y_+$. In particular, this yields the formulas $y^{\phi'} \mapsto T_{y^{\phi'}} = T_0 T_{s_{\phi'}}$, $y^{\lambda_\pi} \mapsto T_{y^{\lambda_\pi}} = \pi T_{v_\pi^{-1}}$.

One verifies using the alcove picture that the elements $\phi(y^\lambda)$ and the generators $T_i$ of $\mathcal{B}(W_0)$ satisfy the defining relations of $\mathcal{B}(Y, W_0)$. Hence $\phi$ extends to a homomorphism $\mathcal{B}(Y, W_0) \to \mathcal{B}(W_e)$. Next one verifies (with the help of Lemma 4.20, below) that the element $y^{\phi'} T_{s_{\phi'}}^{-1} \in \mathcal{B}(Q_0', W_0)$ satisfies braid relations with the generators $T_i$, giving a homomorphism $\mathcal{B}(W_a) \to \mathcal{B}(Q_0', W_0)$ inverse to $\phi$. Hence $\phi$ maps $\mathcal{B}(Q_0', W_0)$ isomorphically onto $\mathcal{B}(W_a)$, and by Theorem 2.11, it follows that $\phi$ is an isomorphism.

**Corollary 4.3.** *For a (left) extended affine Weyl group $W_a \rtimes \Pi = W_0 \ltimes X$, there is an isomorphism $\mathcal{B}(W_e) \cong \mathcal{B}(W_0, X)$ between the extended affine braid group and the left affine braid group of $X$, which is the identity on $\mathcal{B}(W_0)$, and satisfies $x^{-\phi} \leftrightarrow T_{s_\phi} T_0$, $x^{-\lambda_\pi} \leftrightarrow T_{v_\pi} \pi^{-1}$.*

**4.4.** We come now to the key construction in the theory. Fix two finite root systems $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$, $\big(Y, (\alpha_i'), (\alpha_i'^\vee)\big)$ with the same Weyl group $W_0$. More accurately, assume given an isomorphism of Coxeter groups $W_0 = (W(X), S) \cong (W(Y), S')$, and label the simple roots so that $s_i$ corresponds to $s_i'$ for each $i = 1, \dots, n$.

Let $\phi \in Q_0 \subseteq X, \phi' \in Q_0' \subseteq Y$ be the dominant short roots. Let $\theta \in Q_0, \theta' \in Q_0'$ be the dominant roots such that $s_\theta = s_{\phi'}$, $s_{\theta'} = s_\phi$. There are unique $W_0$-equivariant pairings $(X, Q_0') \to \mathbb{Z}$, $(Q_0, Y)' \to \mathbb{Z}$ such that $(\beta, \phi') = \langle \beta, \theta^\vee \rangle$ for all $\beta \in X$ and $(\phi, \beta')' = \langle \beta', \theta'^\vee \rangle$ for all $\beta' \in Y$. One checks that $(\phi, \phi') = (\phi, \phi')' = 2$ if $s_\phi = s_{\phi'}$, and $(\phi, \phi') = (\phi, \phi')' = 1$ if $s_\phi \neq s_{\phi'}$. By $W_0$-equivariance, the two pairings therefore agree on $Q_0 \times Q_0'$. Fix a $W_0$-invariant pairing $(X, Y) \to \mathbb{Q}$ extending the two pairings $(\cdot\,, \cdot)$ and $(\cdot\,, \cdot)'$, and choose $m$ such that $(X, Y) \subseteq \mathbb{Z}/m$.

**Remark 4.5.** The Cartan matrices of $X$ and $Y$ are clearly either of the same type $(Z_n, Z_n)$, or of dual types $(Z_n, Z_n^\vee)$. In the symmetric case $(Z_n, Z_n)$, the roots $\theta = \phi$, $\theta' = \phi'$ are short, and the pairing $(\cdot\,, \cdot)$ restricts on $Q_0 = Q_0'$ to the $W_0$-equivariant pairing such that $(\alpha, \alpha) = 2$ for short roots $\alpha$. In the dual case $(Z_n, Z_n^\vee)$, $\theta$ and $\theta'$ are long, and the pairing restricts to the canonical pairing between $Q_0$ and $Q_0' = Q_0^\vee$. Types $G_2$ and $F_4$ are isomorphic to their duals, but only after relabelling the simple roots. Thus there is a genuine difference between types $(G_2, G_2)$ and $(G_2, G_2^\vee)$, for instance. In particular, $\theta = \phi$ in the first case, and $\theta \neq \phi$ in the second.

**4.6.** Given the data in §4.4, set $\widetilde{X} = X \oplus \mathbb{Z}\delta/m$, $\widetilde{Y} = Y \oplus \mathbb{Z}\delta'/m$. Extend the linear functionals $\alpha_i^\vee$ on $X$ to $\widetilde{X}$ so that $\langle \delta, \alpha_i^\vee \rangle = 0$. Define $\alpha_0 = \delta - \theta$, and let $\alpha_0^\vee$ be the extension of $-\theta^\vee$ such that $\langle \delta, \alpha_0^\vee \rangle = 0$. Making similar definitions in $\widetilde{Y}$, we get two affine root systems

$$\big(\widetilde{X}, (\alpha_0, \dots, \alpha_n), (\alpha_0^\vee, \dots, \alpha_n^\vee)\big), \quad \big(\widetilde{Y}, (\alpha_0', \dots, \alpha_n'), (\alpha_0'^\vee, \dots, \alpha_n'^\vee)\big).$$

Let $Y$ act on $\widetilde{X}$ and $X$ on $\widetilde{Y}$ by

$$y^\lambda(\mu) = \mu - (\mu, \lambda)\delta, \quad x^\mu(\lambda) = \lambda - (\mu, \lambda)\delta'.$$

Since $(\cdot\,,\cdot)$ is $W_0$-invariant, this extends to actions of the extended affine Weyl groups

$$W_e = Y \rtimes W_0, \quad W_e' = W_0 \ltimes X$$

on $\widetilde{X}$ and $\widetilde{Y}$, respectively. The semidirect products $W_e \ltimes \widetilde{X}$, $\widetilde{Y} \rtimes W_e'$ are the (left, right) *extended double affine Weyl groups*. We have the following easy counterpart of Corollary 3.11.

**Corollary 4.7.** *There is a canonical isomorphism $W_e \ltimes \widetilde{X} \cong \widetilde{Y} \rtimes W_e'$, which is the identity on $X$, $Y$ and $W_0$, and maps $q = x^\delta$ to $y^{-\delta'}$. In fact, both groups are identified with $W_0 \ltimes (X \star Y)$, where $X \star Y$ is the Heisenberg group generated by $X$, $Y$ and central element $q^{1/m}$, with relations*

$$x^\mu y^\lambda = q^{(\mu,\lambda)} y^\lambda x^\mu.$$

**Remarks 4.8.** (a) For consistency, set $q = y^{-\delta'}$ in the "right" double affine Weyl group $\widetilde{Y} \rtimes W_e'$. Then the isomorphism maps $q$ to $q$.

(b) When $X$ and $Y$ are of dual types, the affine root systems $\widetilde{X}$, $\widetilde{Y}$ are of untwisted type (§3.6). When $X$ and $Y$ are of the same type, then $\widetilde{X}$, $\widetilde{Y}$ are of dual untwisted type (§3.8).

(c) The requirement that $(\cdot\,,\cdot)$ extend the pairings $(X, Q_0') \to \mathbb{Z}$ and $(Q_0, Y')' \to \mathbb{Z}$ in §4.4 ensures that

$$W_e \ni y^{\phi'} s_{\phi'} = s_0 \in W(\widetilde{X}), \quad W_e' \ni s_\phi x^{-\phi} = s_0' \in W(\widetilde{Y}).$$

Under the action of $W_e = Y \rtimes W_0 = \Pi \ltimes W_a$ on $\widetilde{X}$, the subgroup $W_a = Q_0' \rtimes W_0$ is therefore identified with the Weyl group of $\widetilde{X}$. By Corollary 3.11, $\Pi \subset W_e$ acts on $\widetilde{X}$ by automorphisms of the root system, *i.e.* it permutes the affine simple roots and coroots. So $W_e$ acts on $\widetilde{X}$ as the semi-direct product of the Weyl group $W_a$ and the group of automorphisms $\Pi$. In particular, the extended double affine Weyl group $W_e \ltimes \widetilde{X}$ is the semidirect product

$$\Pi \ltimes (W_a \ltimes \widetilde{X})$$

of $\Pi$ with the affine Weyl group (§2.8) of the affine root system $\widetilde{X}$. Similar remarks apply to $\widetilde{Y} \rtimes W_e'$.

**4.9.** Since $\Pi$ acts by automorphisms of the affine root system $\widetilde{X}$, it also acts naturally on $\mathcal{B}(W_a, \widetilde{X})$ (§2.9), and we can form the semidirect product $\Pi \ltimes \mathcal{B}(W_a, \widetilde{X})$, which we may regard as an extended (left) affine braid group $\mathcal{B}(W_e, \widetilde{X})$ of the affine root system $\widetilde{X}$. Similarly, we can define $\mathcal{B}(\widetilde{Y}, W_e') = \mathcal{B}(\widetilde{Y}, W_a') \rtimes \Pi'$. Define $q = x^\delta$ in $\mathcal{B}(W_e, \widetilde{X})$, and $q = y^{-\delta'} \in \mathcal{B}(\widetilde{Y}, W_e)$, as in Remark 4.8(a). We come now to the fundamental theorem.

**Theorem 4.10.** *The isomorphism $W_e \ltimes \widetilde{X} \cong \widetilde{Y} \rtimes W_e'$ lifts to an isomorphism $\mathcal{B}(W_e, \widetilde{X}) \cong \mathcal{B}(\widetilde{Y}, W_e')$, which is the identity on $X$, $Y$, and $\mathcal{B}(W_0)$, and maps*

$q = x^\delta$ *to* $q = y^{-\delta'}$. (*Here X, Y are identified with their images under* $\mathcal{B}(W_0, X) \cong \mathcal{B}(W_e') \to \mathcal{B}(\widetilde{Y}, W_e')$ *and* $\mathcal{B}(Y, W_0) \cong \mathcal{B}(W_e) \to \mathcal{B}(W_e, \widetilde{X})$, *using Theorem* 4.2 *and Corollary* 4.3)

The group $\mathcal{B}(W_e, \widetilde{X}) = \mathcal{B}(\widetilde{Y}, W_e')$ is the (extended) *double affine braid group*.

**4.11.** By §2.9, there is an isomorphism $\Phi\colon \mathcal{B}(\widetilde{Y}, W_e') \to \mathcal{B}(W_e', \widetilde{Y})$ given by

$$\begin{aligned} &\Phi(y^\lambda) = y^\lambda, && \Phi(T_0') = T_0'^{-1}, \\ &\Phi(\pi) = \pi' \quad (\pi' \in \Pi'), && \Phi(T_i) = T_i^{-1} \quad (i = 1, \dots, n). \end{aligned}$$

The element $T_0$ in $\mathcal{B}(\widetilde{Y}, W_e')$ is defined by $T_0 = y^{\phi'} T_{s_{\phi'}}^{-1}$, whereas $T_0$ in $\mathcal{B}(W_e', \widetilde{Y})$ is given by $T_0 = T_{s_{\phi'}}^{-1} y^{-\phi'}$. Similarly, $\Pi \to \mathcal{B}(\widetilde{Y}, W_e')$ is given by $\pi = y^{\lambda_\pi} v_\pi \mapsto y^{\lambda_\pi} T_{v_\pi^{-1}}^{-1}$, whereas $\Pi \to \mathcal{B}(W_e', \widetilde{Y})$ is given by $\pi^{-1} \mapsto T_{v_\pi}^{-1} y^{-\lambda_\pi}$, and therefore $\pi \mapsto y^{\lambda_\pi} T_{v_\pi}$. Moreover, $X$ is embedded in $\mathcal{B}(\widetilde{Y}, W_e')$ via the identification $\mathcal{B}(W_e') = \Pi' \ltimes \mathcal{B}(W_0, X)$, which is characterized by $x^{-\phi} \mapsto T_{s_\phi} T_0$ and $x^{-\lambda_{\pi'}} \mapsto T_{v_{\pi'}} \pi'^{-1}$, whereas $X \subset \mathcal{B}(W_e', \widetilde{Y})$ is given via $\mathcal{B}(W_e') = \mathcal{B}(X, W_0) \rtimes \Pi'$ by $x^\phi \mapsto T_0 T_{s_\phi}$, $x^{\lambda_{\pi'}} \mapsto \pi' T_{v_{\pi'}^{-1}}$. In $\mathcal{B}(W_e', \widetilde{Y})$, finally, $q$ denotes $y^{\delta'}$. Taking into account that $\Phi(T_w) = T_{w^{-1}}^{-1}$ for all $w \in W_0$, all this implies

$$\begin{aligned} &\Phi(x^\mu) = x^\mu, && \Phi(T_0) = T_0^{-1}, \\ &\Phi(\pi) = \pi \quad (\pi \in \Pi), && \Phi(q) = q^{-1}. \end{aligned}$$

Theorem 4.10 therefore has the following equivalent alternate formulation.

**Corollary 4.12.** *There is an isomorphism* $\mathcal{B}(W_e, \widetilde{X}) \cong \mathcal{B}(W_e', \widetilde{Y})$, *which is the identity on X, Y,* $\Pi$ *and* $\Pi'$, *maps* $q = x^\delta$ *to* $q^{-1} = y^{-\delta'}$, *and maps the generators* $T_i$ *of* $\mathcal{B}(W_0)$ *to* $T_i^{-1}$.

**4.13.** Cherednik [1] announced Theorem 4.10 in the case $X = Y$, and suggested a possible topological proof, which was completed by Ion [7]. Macdonald [16, 3.5–3.7] gave an elementary proof, which however involves quite a bit of case-checking and only applies when $X = \mathrm{Hom}(Q_0^\vee, \mathbb{Z})$, $Y = \mathrm{Hom}(Q_0'^\vee, \mathbb{Z})$. We now outline a different elementary proof. First assume that the theorem holds in the "unextended" case, $X = Q_0$, $Y = Q_0'$, $W_e = W_a$, $W_e' = W_a'$. We will deduce the general case.

By Theorem 2.11, $\mathcal{B}(\widetilde{Q}_0', W_a')$ embeds in $\mathcal{B}(\widetilde{Y}, W_a')$ as a normal subgroup, with quotient $\widetilde{Y}/\widetilde{Q}_0' = Y/Q_0' \cong \Pi$. Moreover, $\Pi \subseteq \mathcal{B}(W_e) = \mathcal{B}(Y, W_0)$ is a subgroup of $\mathcal{B}(\widetilde{Y}, W_a')$, giving the semidirect decomposition $\mathcal{B}(\widetilde{Y}, W_a') \cong \Pi \ltimes \mathcal{B}(\widetilde{Q}_0', W_a')$. By assumption, we have $\mathcal{B}(W_a, \widetilde{Q}_0) \cong \mathcal{B}(\widetilde{Q}_0', W_a')$, hence $\mathcal{B}(\widetilde{Y}, W_a') \cong \Pi \ltimes \mathcal{B}(W_a, \widetilde{Q}_0) = \mathcal{B}(W_e, Q_0)$. This establishes the case where $X = Q_0$ and $Y$ is general. Exchanging $X$ and $Y$, we also get the case $Y = Q_0'$, $W_e = W_a$, where now $X$ and $W_e'$ are general.

By definition, $\mathcal{B}(\widetilde{Q}_0', W_e') = \mathcal{B}(\widetilde{Q}_0', W_a') \rtimes \Pi'$ and $\mathcal{B}(\widetilde{Y}, W_e') = \mathcal{B}(\widetilde{Y}, W_a') \rtimes \Pi'$, with $\Pi' \cong X/Q_0$ the same for both groups. Again, Theorem 2.11 implies that the first group is a normal subgroup of the second, with quotient $\Pi$. So we can repeat the preceding argument to get the general case.

**4.14.** Now fix $X = Q_0$, $Y = Q_0'$, so $W_e = W_a$, $W_e' = W_a'$. Using Theorem 4.2 and Corollary 4.3, we identify $\mathcal{B}(W_a') = \mathcal{B}(W_0, X)$, $\mathcal{B}(W_a) = \mathcal{B}(Y, W_0)$. Then each group $\mathcal{B}(W_a, \widetilde{X})$, $\mathcal{B}(\widetilde{Y}, W_a')$ has generators $T_0, T_0', T_1, \dots, T_n, q^{1/m}$. In both groups, $q^{1/m}$ is central, the generators $T_0, T_1, \dots, T_n$ satisfy the braid relations of $\mathcal{B}(W_a)$, and $T_0', T_1, \dots, T_n$ satisfy those of $\mathcal{B}(W_a')$.

The additional relations (7) for $\lambda \in Q_0$ and $i = 0$ complete a presentation of $\mathcal{B}(W_a, \widetilde{X})$, since those for $i \neq 0$ already hold in $\mathcal{B}(W_0, X) = \mathcal{B}(W_a')$. For convenience, we write down these extra relations again here, after applying the identity $\langle \lambda, \alpha_0^\vee \rangle = -\langle \lambda, \theta^\vee \rangle$:

$$T_0^a x^\lambda T_0^b = x^{s_0(\lambda)}, \quad \text{where } a, b \in \{\pm 1\} \text{ and } -\langle \lambda, \theta^\vee \rangle = (a+b)/2. \tag{16}$$

In view of Corollary 4.12, to prove the theorem it suffices to express (16) in a "self-dual" form, in the sense that the substitutions $T_0 \leftrightarrow T_0'^{-1}$, $T_i \leftrightarrow T_i^{-1}$, $q \leftrightarrow q^{-1}$ ($i \neq 0$) should transform (16) into its counterpart with the roles of $X$ and $Y$ interchanged.

**Lemma 4.15.** *Relations* (16) *reduce to the case when $\lambda$ is a short positive root $\alpha \neq \theta$ (i.e., $\alpha \neq \phi$ if $\theta = \phi$ is short).*

*Proof.* The short roots $\beta \neq \pm\theta$ span $Q_0$. Hence we can always write $\lambda = \beta_1 + \cdots + \beta_m$, where $\beta_i \in (R_0)_{\text{short}} \setminus \{\pm\theta\}$. In particular, $\langle \beta_i, \theta^\vee \rangle \in \{0, \pm 1\}$ for all $i$. Given that $\langle \lambda, \theta^\vee \rangle \in \{0, \pm 1\}$, we can always order the $\beta_i$ so that those with $\langle \beta_i, \theta^\vee \rangle = 1$ and those with $\langle \beta_i, \theta^\vee \rangle = -1$ alternate. Writing (16) in the form $T_0^a x^\lambda = x^{s_0(\lambda)} T_0^{-b}$, it is easy to see that it follows from the same relation for each $\beta_i$. This reduces us to the case that $\alpha \neq \pm\theta$ is a short root. The case of (16) for $\langle \lambda, \theta^\vee \rangle = 1$ implies the case for $\langle \lambda, \theta^\vee \rangle = -1$, so positive roots $\alpha$ suffice. □

**4.16.** A *parabolic subgroup* of $W_0$ is a subgroup of the form $W_J = \langle s_i : i \in J \rangle$, where $J \subseteq \{1, \dots, n\}$. Since $\phi$ and $\phi'$ are dominant, their stabilizers are parabolic subgroups $W_J$, $W_{J'}$ respectively, where $J = \{i : \langle \phi, \alpha_i^\vee \rangle = 0\}$, and $J' = \{i : \langle \phi', \alpha_i'^\vee \rangle = 0\}$. Recall that each left, right and double coset $vW_J$, $W_{J'}v$, $W_{J'}vW_J$ has a unique representative of minimal length, which is also minimal in the Bruhat order.

**Proposition 4.17.** *Relations* (16) *for $\lambda = \alpha \neq \theta$ a short positive root reduce to relations of the following two forms:*

(a) *For $v$ such that $(v(\phi), \phi') = 0$ and $v$ minimal in $W_{J'}vW_J$, the relation*

$$T_0 T_v T_0'^{-1} T_v^{-1} = T_v T_0'^{-1} T_v^{-1} T_0.$$

(b) *For $v = v_1$ such that $(v(\phi), \phi') = 1$ and $v$ minimal in $W_{J'}vW_J$, define $v_2$, $v_3$, $v_4$ minimal respectively in $W_{J'}vs_\phi W_J$, $W_{J'}s_\theta vs_\phi W_J$, $W_{J'}s_\theta vW_J$; this given, the relation*

$$T_0^{-1}T_{v_1}T_0'^{-1}T_{v_2}^{-1}T_0^{-1}T_{v_3}T_0'^{-1}T_{v_4}^{-1} = q.$$

*Proof.* We can always write $\alpha = v(\phi)$ with $v$ minimal in $vW_J$. If $i \in J'$, then $T_i$ commutes with $T_0$. In $\mathcal{B}(W_a') = \mathcal{B}(W_0, Q_0)$ we have $x^{s_i(\alpha)} = T_i^{\epsilon} x^\alpha T_i^{\epsilon'}$, $\epsilon, \epsilon' = \pm 1$ for every positive short root $\alpha$. These facts imply that relations (16) are invariant under replacement of $\alpha$ with $w(\alpha) \in W_{J'}\alpha$. Hence we can assume $v$ minimal in $W_{J'}vW_J$.

We show that when $\langle \alpha, \theta^\vee \rangle = (v(\phi), \phi') = 0$, relation (16), which in this case reads $T_0 x^\alpha = x^\alpha T_0$, is equivalent to (a). The minimality of $v$ in $vW_J$ implies that if $v = s_{i_1} \dots s_{i_l}$ is a reduced factorization, then $\langle s_{i_{k+1}} \dots s_{i_l}(\phi), \alpha_{i_k}^\vee \rangle = 1$ for all $k$. Hence $x^\alpha = T_v x^\phi T_{v^{-1}} = T_v T_0'^{-1} T_{s_\phi}^{-1} T_{v^{-1}}$. The minimality also implies that $s_\phi = v^{-1}s_\alpha v$ is a reduced factorization. Therefore $T_{s_\phi}^{-1}T_{v^{-1}} = T_v^{-1}T_{s_\alpha}^{-1}$, and $x^\alpha = T_v T_0'^{-1} T_v^{-1} T_{s_\alpha}^{-1}$. Now, since $\langle \alpha, \theta^\vee \rangle = 0$, we have $s_0 s_\alpha = s_\alpha s_0$, and both sides of this equation are reduced factorizations. Hence $T_0$ commutes with $T_{s_\alpha}$, so (16) is equivalent to $T_0$ commuting with $T_v T_0'^{-1} T_v^{-1}$.

For $\langle \alpha, \theta^\vee \rangle = (v(\phi), \phi') = 1$, we have $s_0(\alpha) = \alpha + \alpha_0 = \alpha - \theta + \delta$, and thus relation (16) in this case reads $T_0^{-1} x^\alpha T_0^{-1} = q x^{-\beta}$, or $T_0^{-1} x^\alpha T_0^{-1} x^\beta = q$, where $\beta = -s_\theta(\alpha)$ satisfies $\alpha + \beta = \theta$. Let $u$ be the minimal representative of $s_\theta v s_\phi W_J$. Then $\beta = u(\phi)$, and the same reasoning as in the previous paragraph gives $x^\alpha = T_v T_0'^{-1} T_{vs_\phi}^{-1}$, $x^\beta = T_u T_0'^{-1} T_{us_\phi}^{-1}$. Our relation now takes the form

$$T_0^{-1}T_v T_0'^{-1} T_{vs_\phi}^{-1} T_0^{-1} T_u T_0'^{-1} T_{us_\phi}^{-1} = q. \tag{17}$$

Using §2.6 and the fact that $s_\phi(\alpha_i) = \alpha_i$ for all $i \in J$, we deduce (for any $J'$)

(*) *if $x$, $y$ are minimal in $W_{J'}x$, $W_{J'}y = W_{J'}xs_\phi$, respectively, and $xw$ is minimal in $W_{J'}xW_J$, then $yw$ is minimal in $W_{J'}yW_J$.*

By construction, $u$ and $v$ are minimal in their left $W_J$ cosets, and (*) implies the same for $us_\phi$ and $vs_\phi$. Hence the elements $v_1 = v$, $v_2$, $v_3$, $v_4$ defined in (b) are the minimal representatives of $W_{J'}v$, $W_{J'}vs_\phi$, $W_{J'}u$, $W_{J'}us_\phi$ respectively. By the analog of (*) for $s_\theta$ (operating on the left), we see that $v_1 = v$ implies $v_4 = us_\phi$, and if we set $v_2 = wvs_\phi$, then $v_3 = wu$. Now $w \in W_{J'}$ commutes with $T_0$, and the factorizations $vs_\phi = w^{-1}v_2$, $u = w^{-1}v_3$ are reduced, so (17) reduces to (b). □

**Corollary 4.18.** *The (unextended) double affine braid group $\mathcal{B}(W_a, \widetilde{Q}_0)$, where $\mathcal{B}(W_a) = \mathcal{B}(Q_0', W_0)$, has a presentation with generators $T_0$, $T_0'$, $T_1, \dots, T_n$, $q^{1/m}$ and the following (manifestly self-dual) relations: $q^{1/m}$ is central; braid relations for $T_0, T_1, \dots, T_n \in \mathcal{B}(W_a)$ and for $T_0', T_1, \dots, T_n \in \mathcal{B}(W_a')$; and the relations in Proposition* 4.17.

**Example 4.19.** Let $X = Y$ be of type $A_{n-1}$, with $\alpha_i = \alpha_i^\vee = e_i - e_{i+1}$ as in Example 2.5. Then $\phi = \theta = \phi' = \theta' = e_1 - e_n$, and $W_J = W_{J'} = \langle s_2, \dots, s_{n-2}\rangle$. The presentation of $\mathcal{B}(W_a, \widetilde{Q}_0)$ is given by $q$ central, braid relations and

(a) $T_0$ commutes with $T_1 T_{n-1} T_0'^{-1} (T_1 T_{n-1})^{-1}$,

(b) $T_0^{-1} T_1 T_0'^{-1} T_1^{-1} T_2^{-1} \dots T_{n-1}^{-1} T_0^{-1} T_{n-1} T_0'^{-1} T_{n-1}^{-1} T_{n-2}^{-1} \dots T_1^{-1} = q$.

There are seven double cosets $W_J v W_{J'}$. Two have $v(\phi) = \pm\phi$, one yields (a), and the other four provide the elements $v_1, \dots, v_4$ in (b). In fact, in every type there turns out to be only one relation of type (b) and at most two of type (a), except for $\widetilde{D}_4$, which has three of type (a).

**Lemma 4.20.** *If $\phi$ is the dominant short root of a finite root system $X$, and $v \in W_0$ is such that $\alpha = v(\phi) \in (R_0)_+$, then in $\mathcal{B}(W_0, X)$ we have*

$$T_v x^\phi T_{s_\phi} T_v^{-1} = x^\alpha T_{s_\alpha}.$$

*Proof.* This reduces to the case that $v$ is minimal in $vW_J$ (in the notation of §4.16). As in the proof of Proposition 4.17 we then have $T_v x^\phi T_{s_\phi} = x^\alpha T_{v^{-1}}^{-1} T_{s_\phi} = x^\alpha T_{s_\alpha} T_v$. □

Lemma 4.20 will be used in the proof of Theorem 5.11. Its variant for $\mathcal{B}(Y, W_0)$ is $T_{v^{-1}}^{-1} y^\phi T_{s_\phi}^{-1} T_{v^{-1}} = y^\alpha T_{s_\alpha}^{-1}$, which is useful for verifying the braid relations in the proof of Theorem 4.2.

## 5. Hecke algebras and Cherednik algebras

**5.1.** Let $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$ be a non-degenerate root system, with Cartan matrix $A$, Weyl group $W$, and roots $R$. To each $W$-orbit in $R$ we associate a parameter $u_\alpha$, $u_\alpha = u_\beta$ if $\beta = w(\alpha)$. Set $u_i = u_{\alpha_i}$. The $u_i$ are assumed to be invertible elements of some commutative ground ring $\mathfrak{A}$. If $\alpha_i^\vee \in 2X^\vee$, we also introduce a second parameter $u_i'$.

**Lemma 5.2.** *Let $\mathcal{H}$ be an $\mathfrak{A}$-algebra containing the group algebra $\mathfrak{A}X$, and $T_i \in \mathcal{H}$.*

(i) *If $\alpha_i^\vee \notin 2X^\vee$, then commutation relations* (5)–(6) *and the quadratic relation*

$$(T_i - u_i)(T_i + u_i^{-1}) = 0 \tag{18}$$

*imply the more general commutation relations, for all $\lambda \in X$,*

$$T_i x^\lambda - x^{s_i(\lambda)} T_i = \frac{(u_i - u_i^{-1})}{1 - x^{\alpha_i}} (x^\lambda - x^{s_i(\lambda)}). \tag{19}$$

(ii) *If $\alpha_i^\vee \in 2X^\vee$, then* (5)–(6), (18) *and the additional quadratic relation*

$$(T_i^{-1} x^{-\alpha_i} - u_i')(T_i^{-1} x^{-\alpha_i} + u_i'^{-1}) \tag{20}$$

*imply*

$$T_i x^\lambda - x^{s_i(\lambda)} T_i = \frac{(u_i - u_i^{-1}) + (u_i' - u_i'^{-1}) x^{\alpha_i}}{1 - x^{2\alpha_i}} (x^\lambda - x^{s_i(\lambda)}) \tag{21}$$

(iii) *Given* (18)*, relation* (21) *implies* (20)*, and* (19) *implies that* (20) *holds with* $u_i' = u_i$.

Note that the denominators in (19), (21) divide $x^\lambda - x^{s_i(\lambda_i)}$.

For the well-known proof, observe that each side of (19), (21), viewed as an operator on $x^\lambda$, satisfies $F(x^\lambda x^\mu) = F(x^\lambda)x^\mu + x^{s_i(\lambda)} F(x^\mu)$. Hence (19), (21) for $x^\lambda, x^\mu$, imply the same for $x^{\lambda \pm \mu}$. This reduces (i) to the special cases $\langle \lambda, \alpha_i^\vee \rangle \in \{0, 1\}$, which in turn reduce to (5)–(6), using the identity $T_i^{-1} = T_i - u_i + u_i^{-1}$, which is equivalent to (18). Similarly, (ii) reduces to the special cases $\langle \lambda, \alpha_i^\vee \rangle = 0$, which is (5) ((6) is vacuous if $\alpha_i^\vee \in 2X^\vee$), and $\lambda = \alpha_i$ (since $\langle \alpha_i, \alpha_i^\vee \rangle = 2$). Modulo (18), this last case is equivalent to (20), which also gives (iii) in case (ii). For (iii) in case (i), observe that (19) is just (21) with $u_i' = u_i$.

**Definition 5.3.** The *affine Hecke algebra* $\mathcal{H}(W, X)$ is the quotient $(\mathfrak{A}\mathcal{B}(W, X))/\mathfrak{j}$, where $\mathfrak{j}$ is the 2-sided ideal generated by the quadratic relations (18) for all $i$, plus (20) for each $i$ such that $\alpha_i^\vee \in 2X^\vee$.

Equivalently, $\mathcal{H}(W, X)$ is generated by elements $x^\lambda$ ($\lambda \in X$) and $T_i$ satisfying the braid relations of $\mathcal{B}(W)$, quadratic relations (18), and relations (19) or (21) depending on whether or not $\alpha_i^\vee \in 2X^\vee$.

**Proposition 5.4.** *The subalgebra of* $\mathcal{H}(W, X)$ *generated by the elements* $T_i$ *is isomorphic to the ordinary Hecke algebra* $\mathcal{H}(W)$*, with basis* $\{T_w : w \in W\}$*, and* $\mathcal{H}(W, X)$ *has basis* $\{T_w x^\lambda\}$.

*Proof.* The commutation relations (19), (21) imply that the elements $T_w x^\lambda$ span; they are independent because the specialization $u_i = u_i' = 1$ collapses $\mathcal{H}(W, X)$ to the group algebra $\mathfrak{A} \cdot (W \ltimes X)$. (More precisely, specialization implies the result for $\mathfrak{A} = \mathbb{Z}[u_i^{\pm 1}, u_i'^{\pm 1}]$, and the general case follows by extension of scalars.) □

**5.5.** Let $\Pi$ be a group acting by automorphisms of the root system $\big(X, (\alpha_i), (\alpha_i^\vee)\big)$, and assume that $u_i = u_j$, $u_i' = u_j'$ for $\alpha_j \in \Pi(\alpha_i)$. Then $\Pi$ acts on $\mathcal{H}(W, X)$, and we define the *extended affine Hecke algebra* to be the twisted group algebra $\Pi \cdot \mathcal{H}(W, X)$ generated by $\Pi$ and $\mathcal{H}(W, X)$ with relations $\pi f = \pi(f)\pi$ for $\pi \in \Pi$, $f \in \mathcal{H}(W, X)$.

Up to now the root system $X$ was arbitrary. If $X$ is finite, with $W_0 \ltimes X = W_a \rtimes \Pi$ as in Corollary 4.3, then $\mathcal{H}(W_0, X)$ is isomorphic to the twisted group algebra $\mathcal{H}(W_a) \cdot \Pi$ of the ordinary Hecke algebra of $W_a$. The most interesting case is when $X$ is affine; specifically when $X = \widetilde{X}$ as constructed in §4.6.

**Definition 5.6.** Given $X$, $Y$, $(\cdot, \cdot)$, $\widetilde{X}$, $\widetilde{Y}$, $W_e = \Pi \ltimes W_a$, $W_e' = W_a' \rtimes \Pi'$ as in §4.4–4.9, the (left) *Cherednik algebra* $\mathcal{H}(W_e, \widetilde{X})$ is the extended affine Hecke algebra $\Pi \cdot \mathcal{H}(W_a, \widetilde{X})$.

Equivalently, $\mathcal{H}(W_e, \widetilde{X})$ is generated by $x^\lambda \in X$, $\pi \in \Pi$, $T_0, \ldots, T_n$ and $q^{\pm 1/m}$, satisfying the relations of the double affine braid group $\mathcal{B}(W_e, \widetilde{X})$ and the quadratic relations (18), plus (20) if $\alpha_i^\vee \in 2\widetilde{X}^\vee$.

**5.7.** We will also define a *right* Cherednik algebra $\mathcal{H}(\widetilde{Y}, W_e')$, but first we must re-index the parameters. For convenience, we define $u_j' = u_j$ if $\alpha_j^\vee \notin 2\widetilde{X}^\vee$. Define $u_{i'} = u_i$ for $i \neq 0$, and set $u_{0'} = u_j'$, where $\alpha_j$ is a short simple root of the finite root system $X$. If $\alpha_i'^\vee \in 2\widetilde{Y}^\vee$ for $i \neq 0$ (there is at most one such index $i$), set $u_{i'}' = u_0$ . If $\alpha_0'^\vee \in 2\widetilde{Y}^\vee$, set $u_{0'}' = u_0'$.

We now define $\mathcal{H}(\widetilde{Y}, W_e')$ to be the algebra with generators $y^\mu$ ($\mu \in Y$), $\pi' \in \Pi'$, $T_0', T_1, \ldots, T_n, q^{\pm 1/m}$ satisfying the relations of the right affine braid group $\mathcal{B}(\widetilde{Y}, W_e')$, relations (18) with $u_{i'}$ in place of $u_i$, and for $\alpha_i'^\vee \in 2\widetilde{Y}^\vee$, the relations

$$(T_i'^{-1} y^{\alpha_i'} - u_{i'}')(T_i'^{-1} y^{\alpha_i'} + u_{i'}'^{-1}), \tag{22}$$

where we define $T_i' = T_i$ if $i \neq 0$.

**Corollary 5.8.** *The elements* $\{y^\mu T_w x^\lambda\}$ ($\mu \in Y$, $\lambda \in X$, $w \in W_0$) *form an* $\mathfrak{A}[q^{\pm 1/m}]$-*basis of the Cherednik algebras* $\mathcal{H}(W_e, \widetilde{X})$, $\mathcal{H}(\widetilde{Y}, W_e')$.

This follows easily from Proposition 5.4 for $\mathcal{H}(W_e, \widetilde{X})$ and by symmetry for $\mathcal{H}(\widetilde{Y}, W_e')$. We remark that the factors $y^\mu T_w x^\lambda$ can be taken in any order.

**Lemma 5.9.** *We have* $\alpha_0'^\vee \in 2\widetilde{Y}^\vee$ *if and only if* $X$, $Y$ *are both of type* $B_n$ *and* $\Pi$ *acts trivially on the simple roots of* $\widetilde{X}$.

*Proof.* By definition, $\alpha_0'^\vee = -\theta'^\vee$. We can only have $\theta'^\vee \in 2Y^\vee$ if $Y$ is of type $B_n$ and $\theta = \phi$ is short, hence $X$ is also of type $B_n$. Let $P_0'$ be the image of the canonical homomorphism $Y \to \mathrm{Hom}(Q_0'^\vee, \mathbb{Z})$. For type $B_n$ we have either $Q_0' = P_0'$ or $P_0'/Q_0' \cong \mathbb{Z}/2\mathbb{Z}$, with $Q_0' = P_0'$ iff the short roots $\alpha'$ satisfy $\alpha'^\vee \in 2Y^\vee$. The isomorphism $\Pi \cong Y/Q_0'$ (Corollary 3.12) identifies $P_0'/Q_0'$ with the quotient of $\Pi$ by the kernel of its action on the simple roots of $\widetilde{X}$. □

**Remark 5.10.** If $X$, $Y$ are of type $B_n$, then $\widetilde{X}$, $\widetilde{Y}$ are of type $\widetilde{C}_n^\vee$. Label the Dynkin diagram

$$\begin{array}{ccccccccccc} \bullet & \Leftarrow & \bullet & - & \bullet & - & \cdots & - & \bullet & \Rightarrow & \bullet \\ 0 & & 1 & & 2 & & & & n-1 & & n \end{array} \quad . \tag{23}$$

If $\alpha_0'^\vee \notin 2\widetilde{Y}^\vee$, then $\Pi$ acts non-trivially, exchanging nodes 0 and $n$, and similarly for $\alpha_0^\vee$ and $\Pi'$. The four associated parameters are related by the diagram

$$\begin{array}{ccc} (u_0' = u_{0'}') & = & (u_n' = u_{0'}) \\ \| & & \| \\ (u_0 = u_{n'}') & = & (u_n = u_{n'}) \end{array} , \tag{24}$$

where the horizontal equalities hold if $\alpha'^\vee \notin 2Y^\vee$ for short roots $\alpha' \in Y$, and the vertical ones hold if $\alpha^\vee \notin 2X^\vee$ for short roots $\alpha \in X$.

**Theorem 5.11.** *There is an isomorphism $\mathcal{H}(W_e, \widetilde{X}) \cong \mathcal{H}(\widetilde{Y}, W'_e)$, which is the identity on all the generators $X$, $Y$, $q$, $T_i$, $T_0$, $T'_0$, $\pi$, $\pi'$.*

*Proof.* For the most part, this is Theorem 4.10, but we must prove that relations (22) and the case of (18) for $T'_0$ hold in $\mathcal{H}(W_e, \widetilde{X})$. By definition, $T'_0 = T_{s_\phi}^{-1} x^{-\phi}$. By Lemma 4.20, this is conjugate to $T_j^{-1} x^{-\alpha_j}$ for a short simple root $\alpha_j$. Then (20) for $T_j$ implies (18) for $T'_0$. Similarly, if $i \neq 0$ in (22), then $\alpha'_i$ is short, and $T_i'^{-1} y^{\alpha_i}$ is conjugate to $y^{\alpha_i} T_i'^{-1}$ and in turn to $T_0 = y^{\phi'} T_{s_{\phi'}}^{-1}$. By Lemma 5.9, we only have $i = 0$ in (22) when $X$, $Y$ are both of type $B_n$, so $\theta = \phi$, $\theta' = \phi'$. Then $T_0'^{-1} y^{\alpha'_0} = q^{-1} x^{\phi} T_{s_\phi} y^{-\phi'} = x^{-\alpha_0} T_0^{-1}$, which is conjugate to $T_0^{-1} x^{-\alpha_0}$. □

**Corollary 5.12.** *Assume given an automorphism $\varepsilon : \mathfrak{A} \to \mathfrak{A}$ such that $\varepsilon(u_i) = u_i^{-1}$, $\varepsilon(u'_i) = u_i'^{-1}$. Then there is an $\varepsilon$-linear isomorphism $\mathcal{H}(W_e, \widetilde{X}) \cong \mathcal{H}(W'_e, \widetilde{Y})$ which is the identity on $X$, $Y$, $\Pi$, $\Pi'$, maps $q$ to $q^{-1}$, and maps $T_i$ to $T_i^{-1}$ for all $i = 0', 0, 1, \ldots, n$, where the parameters $u_{i'}$, $u'_{i'}$ for $\mathcal{H}(W'_e, \widetilde{Y})$ are as in* §5.7.

*Proof.* The map $\Phi$ in §4.11, composed with $\varepsilon$, preserves (18) and interchanges (22) with the version of (20) for $\widetilde{Y}$ in place of $\widetilde{X}$. □

**5.13.** Let $\mathcal{H} = \Pi \cdot \mathcal{H}(W, X)$ be an extended affine Hecke algebra. The ordinary (extended) Hecke algebra $\Pi \cdot \mathcal{H}(W)$ has a one-dimensional representation $\mathbf{1} = \mathfrak{A}e$ such that $\pi e = e$, $T_i e = u_i e$. The induced representation $\mathrm{Ind}_{\Pi \mathcal{H}(W)}^{\mathcal{H}}(\mathbf{1})$ is the *polynomial representation*. Proposition 5.4 implies that it is isomorphic to the left regular representation $\mathfrak{A}X$ of $X$, with $\Pi$ acting via its action on $X$, and $T_0, \ldots, T_n$ acting as the operators

$$T_i = u_i s_i + \frac{(u_i - u_i^{-1})}{1 - x^{\alpha_i}}(1 - s_i) \tag{25}$$

$$= u_i - u_i^{-1} \frac{1 - u_i^2 x^{\alpha_i}}{1 - x^{\alpha_i}}(1 - s_i) \tag{26}$$

$$= -u_i^{-1} + u_i(1 + s_i) \frac{1 - u_i^{-2} x^{\alpha_i}}{1 - x^{\alpha_i}} \tag{27}$$

or, if $\alpha_i^\vee \in 2X^\vee$,

$$T_i = u_i s_i + \frac{(u_i - u_i^{-1}) + (u'_i - u_i'^{-1}) x^{\alpha_i}}{1 - x^{2\alpha_i}}(1 - s_i) \tag{28}$$

$$= u_i - u_i^{-1} \frac{(1 - u_i u'_i x^{\alpha_i})(1 + (u_i/u'_i) x^{\alpha_i})}{1 - x^{2\alpha_i}}(1 - s_i) \tag{29}$$

$$= -u_i^{-1} + u_i(1 + s_i) \frac{(1 - (u_i u'_i)^{-1} x^{\alpha_i})(1 + (u'_i/u_i) x^{\alpha_i})}{1 - x^{2\alpha_i}}. \tag{30}$$

In particular, these operators satisfy braid relations. The quadratic relations can be seen directly from (26)–(27) and (29)–(30). The polynomial representation specializes at $u_i = u_i' = 1$ to the $\mathfrak{A}$-linearization of the action of $\Pi \ltimes (W \ltimes X)$ on $X$. It is faithful if $\Pi$ acts faithfully.

**5.14.** For any root $\alpha \in R$, define a partial ordering on $X$ by $\mu <_\alpha \lambda$ if $\lambda - \mu \in \mathbb{Z}\alpha$ and $|\langle \mu, \alpha^\vee \rangle| < |\langle \lambda, \alpha^\vee \rangle|$, or $\langle \mu, \alpha^\vee \rangle = -\langle \lambda, \alpha^\vee \rangle > 0$. Each *root string* $\lambda + \mathbb{Z}\alpha$ is totally ordered by $<_\alpha$. Explicitly,

$$\begin{aligned} &\lambda <_\alpha \lambda + \alpha <_\alpha \lambda - \alpha <_\alpha \lambda + 2\alpha <_\alpha \lambda - 2\alpha <_\alpha \cdots \quad \text{if } \langle \lambda, \alpha^\vee \rangle = 0, \\ &\lambda <_\alpha \lambda - \alpha <_\alpha \lambda + \alpha <_\alpha \lambda - 2\alpha <_\alpha \lambda + 2\alpha <_\alpha \cdots \quad \text{if } \langle \lambda, \alpha^\vee \rangle = 1. \end{aligned}$$

If $B \subseteq R$, define $<_B$ to be the transitive closure of the union $\bigcup_{\alpha \in B} <_\alpha$. In general $<_B$ is not a partial order; we may have $\lambda <_B \lambda$.

**Proposition 5.15.** *Let $w \in W$, $B = R_+ \cap w^{-1}(-R_+)$. In the polynomial representation, we have*

$$T_w(x^\lambda) = u^{\rho_B(\lambda)} x^{w(\lambda)} + \sum_{\mu <_{w(B)} w(\lambda)} a_\mu x^\mu,$$

*where $a_\mu \in \mathfrak{A}$ and $u^{\rho_B(\lambda)} = \prod_{\alpha \in B} u_\alpha^{\sigma(-\langle \lambda, \alpha^\vee \rangle)}$, $\sigma(k) = \pm 1$ as $k \geq 0$ or $k < 0$.*

*Proof.* The case $w = s_i$ follows from formulas (25), (28), and the general case by induction on $l(w)$, using the fact that if $w = s_i v > v$ and $B' = R_+ \cap v^{-1}(-R_+)$, then $B = B' \cup \{v^{-1}(\alpha_i)\}$. □

## 6. Macdonald polynomials

**6.1.** Let $\left(\widetilde{X}, (\alpha_i), (\alpha_i^\vee)\right)$ be a non-degenerate reduced affine root system (§3). As always, we take $i = 0$ to be an affine node, denote the Weyl group, roots, etc. by $W$, $R$, $R_+$, $Q$, $Q_+$, and let $W_0$, $R_0$, $Q_0$, etc. denote the same for the finite root system with simple roots $\alpha_1, \ldots, \alpha_n$. We also allow non-reduced affine root systems, regarded as extensions (§2.12) of a reduced affine root system $\widetilde{X}$, with a larger set of roots $R$. In the non-reduced case, we do not give the extra simple roots their own symbols, but designate them simply as $2\alpha_i$.

Let $\delta$ be the nullroot, and assume that the dual of $\widetilde{X}$ is degenerate, *i.e.*, $\delta^\vee = 0$. Possibly after adjoining a fractional multiple of $\delta$, we can always assume that $\widetilde{X} = X \oplus \mathbb{Z}\delta/m$, where $Q_0 \subseteq X$. Fix such a decomposition.

To each $i$ such that $2\alpha_i \notin R$, we associate a parameter $u_i$ and put $t_i = u_i^2$. To each $i$ such that $2\alpha_i \in R$ we associate two parameters $u_i$, $u_i'$ and put $t_i = u_i u_i'$, $t_i' = u_i / u_i'$. We require that simple roots in the same $W$-orbit have the same parameters, and put $t_\alpha = t_{\alpha_i}$, $t_\alpha' = t_{\alpha_i}'$ if $\alpha \in W(\alpha_i)$. We denote by $\mathbb{Q}(t)$ the field of rational functions in the parameters. The group algebra $\mathbb{Q}(t)\widetilde{X}$ is the ring of Laurent polynomials

$\mathbb{Q}(t)[x^{\pm\varepsilon_1}, \ldots, x^{\pm\varepsilon_N}]$, where $\{\varepsilon_1, \ldots, \varepsilon_N\}$ is a basis of $\widetilde{X}$. As in §4, we let $q = x^\delta$. Then $\mathbb{Q}(t)\widetilde{X} = \mathbb{Q}(t)[q^{\pm 1/m}]X$, and we identify it with a subring of $\mathbb{Q}(q,t)X$.

As in §3.7, let $W = Q_0' \rtimes W_0$, where $Q_0' = Q_0^\vee$ if $\widetilde{X}$ is of untwisted type or $\widetilde{BC}_n$, and $Q_0' = Q_0$ otherwise. In either case, $Q_0'$ acts on $\mathbb{Q}(q,t)X$ by the formula

$$y^\mu(x^\lambda) = q^{-(\lambda,\mu)}x^\lambda, \tag{31}$$

in terms of the $W_0$-invariant pairing $(Q_0, Q_0') \to \mathbb{Z}$ in §4.4 (see also §4.6).

**6.2.** Let $\mathbb{Q}(q,t)\widehat{X}$ denote the $\mathbb{Q}(q,t)$-vector space of possibly infinite formal linear combinations $f = \sum_{\lambda\in X} a_\lambda x^\lambda$. The space $\mathbb{Q}(q,t)\widehat{X}$ is a $\mathbb{Q}(q,t)X$-module – *i.e.*, it makes sense to multiply $f \in \mathbb{Q}(q,t)\widehat{X}$ by $p \in \mathbb{Q}(q,t)X$. We regard $\mathbb{Q}(q,t)X$ as a submodule of $\mathbb{Q}(q,t)\widehat{X}$. Write

$$[x^\lambda]f = a_\lambda$$

for the coefficient of $x^\lambda$ in $f$. Let $\overline{\cdot}$ denote the involution on $\mathbb{Q}(q,t)$ and $\mathbb{Q}(q,t)X$ such that

$$\overline{u_i} = u_i^{-1}, \quad \overline{u_i'} = u_i'^{-1}, \quad \overline{t_\alpha} = t_\alpha^{-1}, \quad \overline{t_\alpha'} = t_\alpha'^{-1}, \quad \overline{q} = q^{-1}, \quad \overline{x^\lambda} = x^{-\lambda}.$$

It extends to $\mathbb{Q}(q,t)\widehat{X}$ by the rule $\overline{\sum_\lambda a_\lambda x^\lambda} = \sum_\lambda \overline{a_\lambda} x^{-\lambda}$. The following theorem is due to Cherednik.

**Theorem 6.3.** *There is a unique element $\Delta_0 = \overline{\Delta_0} \in \mathbb{Q}(q,t)\widehat{Q_0} \subseteq \mathbb{Q}(q,t)\widehat{X}$ with constant term $[1]\Delta_0 = 1$, such that for each Coxeter generator $s_i$ of $W$,*

$$s_i(\Delta_0) = \frac{1 - t_i x^{\alpha_i}}{t_i - x^{\alpha_i}}\Delta_0, \quad \textit{or} \quad s_i(\Delta_0) = \frac{(1 - t_i x^{\alpha_i})(1 + t_i' x^{\alpha_i})}{(t_i - x^{\alpha_i})(t_i' + x^{\alpha_i})}\Delta_0, \tag{32}$$

*where the second formula applies if $2\alpha_i \in R$.*

*Proof.* Define a formal series $\Delta \in \mathbb{Q}[\![q,t]\!]\widehat{Q_0}$ by

$$\Delta = \prod_{\substack{\alpha\in R_+ \\ 2\alpha\notin R,\, \alpha\notin 2R}} \frac{1 - x^\alpha}{1 - t_\alpha x^\alpha} \prod_{\substack{\alpha\in R_+ \\ 2\alpha\in R}} \frac{1 - x^{2\alpha}}{(1 - t_\alpha x^\alpha)(1 + t_\alpha' x^\alpha)}.$$

The coefficients $[x^\lambda]\Delta \in \mathbb{Q}[\![q,t]\!]$ are not rational functions. Define $\Delta_0 = \Delta/([1]\Delta)$. Since $s_i$ leaves the set $R_+ \setminus \{\alpha_i, 2\alpha_i\}$ invariant, it follows that $\Delta$ and $\Delta_0$ satisfy (32). By construction, $\Delta_0$ has constant term 1. These conditions can be expressed as a system of linear equations over $\mathbb{Q}(q,t)$ in the coefficients $[x^\lambda]\Delta_0$, which therefore have a solution $\Delta_0'$ with coefficients in $\mathbb{Q}(q,t)$.

Now, $\Delta_0'/\Delta_0$ is $W$-invariant. For $0 \neq \lambda \in Q_0$, choose $\mu \in Q_0'$ such that $(\lambda,\mu) \neq 0$. Then (31) implies that $[x^\lambda](\Delta_0'/\Delta_0) = 0$, *i.e.*, $\Delta_0'/\Delta_0$ is a constant. Hence $\Delta_0' = \Delta_0$, since they both have constant term 1. This shows that $\Delta_0$ has coefficients in $\mathbb{Q}(q,t)$ and is unique. One checks that (32) is $\overline{\cdot}$-invariant, which implies $\Delta_0 = \overline{\Delta_0}$ by uniqueness. □

The *Macdonald constant term identity* [16, (5.8.20)] provides an explicit infinite product expansion for $[1]\Delta$, but it is not practicable to compute the coefficients of $\Delta_0$ directly from the formula $\Delta_0 = \Delta/([1]\Delta)$. A better procedure is to equate the coefficients of $y^{\phi'}(\Delta_0) = s_0 s_{\phi'}(\Delta_0)$, as given by (31) on the one hand, and by (32) on the other. This leads to a recurrence which determines the coefficients.

**Definition 6.4.** *Cherednik's inner product* on $\mathbb{Q}(q,t)X$ is defined by the formula

$$\langle f, g\rangle_0 = [1](f\,\overline{g}\,\Delta_0).$$

It is linear in $f$ and $\overline{\cdot}$-hermitian by Theorem 6.3, *i.e.*, $\langle g, f\rangle_0 = \overline{\langle f, g\rangle_0}$.

**Lemma 6.5.** *Let $B = (R_0)_+$. Under the identification of $X$ with the set $W_e'/W_0$ of minimal left coset representatives in $W_e' = W_0 \ltimes X$, the ordering $<_B$ defined in* §5.14 *coincides with the Bruhat order $<$ in $W_e'$.*

*Proof.* Let $w_\lambda$ be minimal in $x^\lambda W_0$. If $s_\beta w_\lambda < w_\lambda$ for a reflection $s_\beta \in W_a'$, then clearly $w_{s_\beta(\lambda)} < w_\lambda$. The Bruhat order on $W_e'/W_0$ is the transitive closure of these relations. In the alcove picture (§3.10), $s_\beta$ belongs to a root $\beta = \alpha^\vee + k\delta'$ of the affine root system $X^\vee \oplus \mathbb{Z}\delta'$, where we can assume that $\alpha \in (R_0)_+$. The condition $s_\beta w_\lambda < w_\lambda$ means that $h_\beta$ separates $\Lambda_0^\vee + \lambda$ from the dominant alcove $A_0$. This is equivalent to $s_\beta(\lambda) <_\alpha \lambda$, and $<_B$ is by definition the transitive closure of these relations. □

We fix the partial ordering $<_B$ on $X$, with $B = (R_0)_+$, and denote it by $<$.

**Theorem 6.6.** *There is a unique basis $\{E_\lambda : \lambda \in X\}$ of $\mathbb{Q}(q,t)X$ satisfying the orthogonality and triangularity conditions*

(i) $\langle E_\lambda, E_\mu\rangle_0 = 0$ *for* $\lambda \neq \mu$,

(ii) $E_\lambda = x^\lambda + \sum_{\mu<\lambda} c_{\lambda\mu} x^\mu$, $c_{\lambda\mu} \in \mathbb{Q}(q,t)$.

The $E_\lambda$ are the (non-symmetric) *Macdonald polynomials*. Let us review how their existence and other properties are established using Cherednik algebras.

**6.7.** If $\widetilde{X}$ is of untwisted or dual untwisted type, choose $Y$ and $(X, Y) \to \mathbb{Z}/m$ as in §4.4. One can always take $Y = Q_0'$, but other choices may be more convenient – for instance, in type $A_{n-1}$, it is handy to let $X = Y = \mathbb{Z}^n$ be the weight lattice of $\mathrm{GL}_n$ (Example 2.5).

Non-reduced and mixed types are handled as follows. If $2\alpha_i \in R$, the specialization $u_i' = u_i$, hence $t_i' = 1$, collapses $\Delta_0$ and $\langle\cdot\,,\cdot\rangle_0$ to their counterparts for the root system with $2\alpha_i$ omitted. Similarly, specializing $u_i' = 1$ omits $\alpha_i$. The restriction of $<$ to cosets of the (possibly smaller) root lattice $Q_0$ in the resulting root system does not change. It follows that if Macdonald polynomials $E_\lambda$ exist for the original root system, then they specialize at $u_i' = 1$ (resp. $u_i' = u_i$) to $E_\lambda$ for the root system with $\alpha_i$ (resp. $2\alpha_i$) omitted. To be fully correct, we must also show that the coefficients of $E_\lambda$ do not have poles at these specializations. This will follow from Corollary 6.15.

Every affine root system $\widetilde{X}$ of mixed or non-reduced type embeds as above (perhaps after adjoining $\delta/2$) into a root system of one of two maximally non-reduced types: (a) $\widetilde{X}$ of type $\widetilde{C}_n^\vee$ with $2\alpha_0$, $2\alpha_n$ adjoined (indexing the simple roots as in (23)), or (b) $\widetilde{X}$ of type $\widetilde{B}_n$, with $2\alpha_n$ adjoined. For these types, choose $Y$ and $(\cdot\,,\cdot)$ as for $\widetilde{X}$ of *reduced* type $\widetilde{C}_n^\vee$ or $\widetilde{B}_n$, respectively. Specifically $X$, $Y$ are of types $(B_n, B_n)$ in (a), or $(B_n, C_n)$ in (b), and we have $\alpha^\vee \in 2\widetilde{X}^\vee$ for all short roots $\alpha$. In case (a) we also require $Y$ to satisfy $\alpha'^\vee \in 2\widetilde{Y}^\vee$ for short roots $\alpha'$, so as not to force the parameters for $i = 0$ and $i = n$ to coincide (Remark 5.10).

Let $\mathcal{H} = \mathcal{H}(W_e, \widetilde{X})$ be the Cherednik algebra (Definition 5.6) attached to $X$, $Y$, $(\cdot\,,\cdot)$, with ground ring $\mathfrak{A} = \mathbb{Q}(t)$, and parameters $u_i$ equated with those in §6.1, setting $u_i' = u_i$ in the reduced case. We identify $\mathbb{Q}(q,t)X$ with the underlying space of the polynomial representation (§5.13) of $\mathcal{H}$, after extension of scalars from $\mathbb{Q}(t)[q^{\pm 1}]$ to $\mathbb{Q}(q,t)$. Note that in formulas (25)–(30) for $i = 0$, we have $x^{\alpha_0} = qx^{-\theta}$, and $s_0(x^\lambda) = q^{\langle\lambda,\theta^\vee\rangle} s_\theta(x^\lambda)$, where $\delta = \alpha_0 + \theta$.

**Proposition 6.8.** *The operators $T_i$ (§5.13) are unitary with respect to $\langle\cdot\,,\cdot\rangle_0$.*

*Proof.* For any operator $T$, let $T^*$ denote its adjoint, $\langle T^* f, g\rangle_0 = \langle f, Tg\rangle_0$. We are to show that $T_i^* = T_i^{-1} = T_i - u_i + u_i^{-1}$, or equivalently, since $u_i^* = \overline{u_i} = u_i^{-1}$, that

$$(T_i - u_i)^* = (T_i - u_i).$$

From (32), we deduce that

$$s_i^* = \frac{1 - t_i x^{\alpha_i}}{t_i - x^{\alpha_i}} s_i = s_i \frac{t_i - x^{\alpha_i}}{1 - t_i x^{\alpha_i}}$$

if $2\alpha_i \notin R$, or

$$s_i^* = \frac{(1 - t_i x^{\alpha_i})(1 + t_i' x^{\alpha_i})}{(t_i - x^{\alpha_i})(t_i' + x^{\alpha_i})} s_i = s_i \frac{(t_i - x^{\alpha_i})(t_i' + x^{\alpha_i})}{(1 - t_i x^{\alpha_i})(1 + t_i' x^{\alpha_i})}$$

if $2\alpha_i \in R$. The fractions appearing in these expressions are self-adjoint, since $s_i^* s_i$ and $s_i s_i^*$ are self-adjoint. The result now follows easily from (26)–(27) in the first case (where $t_i = u_i^2$), and (29)–(30) in the second (where $t_i = u_i u_i'$, $t_i' = u_i / u_i'$). □

**Proposition 6.9.** *For $i \neq 0$, introduce formal "logarithms" $k_i$, $k_\alpha = k_i$ for $\alpha \in W_0(\alpha_i)$, with the convention that $q^{k_i} = u_i$. Set*

$$\rho^\vee = \sum_{\alpha \in (R_0)_+} k_\alpha \alpha^\vee, \quad \rho'^\vee = \sum_{\alpha \in (R_0)_+} k_\alpha \alpha'^\vee,$$

*where $\alpha' \in (R_0')_+$ is the positive root such that $s_{\alpha'} = s_\alpha$. Then the* Cherednik operators *$y^\mu \in \mathcal{H}$, acting on $\mathbb{Q}(q,t)X$, satisfy*

$$y^\mu(x^\lambda) = q^{-(\lambda,\mu) + \langle \mu, w_\lambda(\rho'^\vee)\rangle} x^\lambda + \sum_{\mu < \lambda} b_{\lambda\mu} x^\mu, \quad b_{\lambda\mu} \in \mathbb{Q}(q,t), \tag{33}$$

*where $w_\lambda$ is the minimal representative of $x^\lambda W_0$ in $W_e'$.*

*Proof.* It suffices to take $\mu \in Y_+$ dominant, so $y^\mu = T_{y^\mu}$. Bear in mind that $\widetilde{X}$ is now a reduced affine root system of untwisted or dual untwisted type (§6.7), not the root system we started with in §6.1. The affine roots are $\alpha + d\mathbb{Z}\delta$ for $\alpha \in R_0$, where $d = (\alpha, \alpha')/2$, both for untwisted types and their duals. We have $(\alpha, \mu) = d\langle \mu, \alpha'^\vee \rangle$ for all $\mu \in Y$.

If $\beta = \alpha + k\delta$ is a root, then $y^\mu(\beta) = \alpha + (k - (\alpha, \mu))\delta$, and the condition $\beta \in B = R_+ \cap y^{-\mu}(-R_+)$ holds if and only if $0 \le k < (\alpha, \mu)$ and $\alpha \in (R_0)_+$. It follows that for any $\alpha \in (R_0)_+$, the number of roots of the form $\alpha + k\delta \in B$ is equal to $\langle \mu, \alpha'^\vee \rangle$. We also have $x^{y^\mu(\lambda)} = q^{-(\lambda,\mu)} x^\lambda$ by (31). The form of (33) now follows from Proposition 5.15, with leading coefficient given by

$$q^{-(\lambda,\mu)} \prod_{\alpha \in (R_0)_+} u_\alpha^{\langle \mu, \alpha'^\vee \rangle \sigma(-\langle \lambda, \alpha^\vee \rangle)}.$$

This is equal to $q^{-(\lambda,\mu)+\langle \mu, w_\lambda(\rho'^\vee)\rangle}$ because $w_\lambda(\rho'^\vee) = \sum_{\alpha \in (R_0)_+} \pm k_\alpha \alpha'^\vee$, with a minus sign if $\alpha^\vee \in w_\lambda(-(R_0^\vee)_+)$, or equivalently, if $\langle \lambda, \alpha^\vee \rangle > 0$ (see the next remark). □

Note that $\rho^\vee$, $\rho'^\vee$ are characterized by $\langle \alpha_i, \rho^\vee \rangle = \langle \alpha_i', \rho'^\vee \rangle = 2k_i$.

**Remark 6.10.** The action of $W_e' = W_0 \ltimes X$ on $Y^\vee$ factors through $W_0$. Hence, $w_\lambda(\rho'^\vee)$ in (33) depends only on the image of $w_\lambda$ in $W_0$, which is the minimal element $v_\lambda$ such that $v_\lambda^{-1}(\lambda) \in -X_+$. A better way to write (33) is as follows. Define $\Lambda_0^\vee \in \widetilde{Y}_\mathbb{Q}^\vee$ by $\Lambda_0^\vee(Y) = 0$, $\langle \delta', \Lambda_0^\vee \rangle = 1$. Let $\eta\colon X \to Y_\mathbb{Q}^\vee$ be the homomorphism induced by the pairing $(X, Y) \to \mathbb{Q}$, that is, $(\lambda, \mu) = \langle \mu, \eta(\lambda) \rangle$. Then $w_\lambda(\Lambda_0^\vee) = \Lambda_0^\vee + \eta(\lambda)$, and (33) takes the form

$$y^\mu(x^\lambda) = q^{-\langle \mu, w_\lambda(\Lambda_0^\vee - \rho'^\vee)\rangle} x^\lambda + \sum_{\mu < \lambda} b_{\lambda\mu} x^\mu, \quad b_{\lambda\mu} \in \mathbb{Q}(q, t), \tag{34}$$

valid for all $\mu \in \widetilde{Y}$.

**Corollary 6.11.** *Theorem 6.6 holds with $E_\lambda \in (\mathbb{Q}(q,t)Q_0)x^\lambda$ determined uniquely as the joint eigenfunction with eigenvalue $q^{-\langle \mu, w_\lambda(\Lambda_0^\vee - \rho'^\vee)\rangle}$ of the operators $y^\mu$, normalized so that $[x^\lambda]E_\lambda = 1$.*

*Proof.* The $y^\mu$ act on $(\mathbb{Q}(q,t)Q_0)x^\lambda$ as commuting, lower-triangular operators without repeated joint eigenvalues. Since the $y^\mu$ are unitary by Proposition 6.8, their joint eigenfunctions $E_\lambda$ are orthogonal. □

**6.12.** Relation (19) can be written $\phi_i x^\lambda = x^{s_i(\lambda)}\phi_i$, where $\phi_i = T_i - (u_i - u_i^{-1})/(1 - x^{\alpha_i})$. By Corollary 5.12, we also have $\psi_i y^\mu = y^{s_i(\mu)}\psi_i$ for $i = 0', 1, \ldots, n$, where $\psi_i = T_i^{-1} - (u_i^{-1} - u_i)/(1 - y^{\alpha_i'}) = T_i - (u_i - u_i^{-1})/(1 - y^{-\alpha_i'})$, and similarly for (21). It is advantageous to use $u_i\psi_i$ instead here. To this end, set

$$\widetilde{T}_i = u_i T_i \quad (i = 1, \ldots, n);$$
$$\widetilde{T}_{0'} = u_{0'} T_{0'} = u_{0'} T_{s_\phi}^{-1} x^{-\phi} = u_j' T_v^{-1} T_j^{-1} x^{-\alpha_j} T_v = t_\phi \widetilde{T}_v^{-1} \widetilde{T}_j^{-1} x^{-\alpha_j} \widetilde{T}_v,$$

where $s_\phi = v^{-1}s_j v$ is a reduced factorization (Lemma 4.20). These operators depend only on the parameters $t_i, t_i'$. The intertwining relations $u_i\psi_i y^\mu = y^{s_i(\mu)}u_i\psi_i$, along with $\pi' y^\mu = y^{\pi'(\mu)}\pi'$ for $\pi \in \Pi'$ imply the following proposition.

**Proposition 6.13.** *If $E_\lambda$ is a joint eigenfunction of the operators $y^\mu$, $\mu \in \widetilde{Y}$ with eigenvalue $q^{\langle\mu,\Lambda\rangle}$, then $\Psi_i(E_\lambda)$ is a joint eigenfunction with eigenvalue $q^{\langle\mu,s_i(\Lambda)\rangle}$, where $i = 0', 1, \ldots, n$, and*

$$\Psi_i = \widetilde{T}_i + \frac{1-t_i}{1-q^{-\langle\alpha_i',\Lambda\rangle}}, \quad \text{or} \quad \Psi_i = \widetilde{T}_i + \frac{1-t_it_i'+(t_i'-t_i)q^{-\langle\alpha_i',\Lambda\rangle}}{1-q^{-2\langle\alpha_i',\Lambda\rangle}},$$

*the second formula applying in case $\alpha_i'^\vee \in 2\widetilde{Y}^\vee$. Similarly, $\pi'(E_\lambda)$ is a joint eigenfunction with eigenvalue $q^{\langle\mu,\pi'(\Lambda)\rangle}$, for any $\pi' \in \Pi'$.*

**Corollary 6.14.** *For $i \neq 0$, if $s_i(\lambda) = \lambda$, then $s_i E_\lambda = E_\lambda$.*

*Proof.* Proposition 6.13 implies that $T_i E_\lambda$ is a scalar multiple of $E_\lambda$, and from the leading coefficient we deduce $T_i E_\lambda = u_i E_\lambda$, which is equivalent to $s_i E_\lambda = E_\lambda$. □

**Corollary 6.15.** *The Macdonald polynomials satisfy the recurrence*

$$E_{v_{\pi'}(\lambda)+\lambda_{\pi'}} = q^{-\langle\lambda_{(\pi'^{-1})},w_\lambda(\rho^\vee)\rangle}x^{\lambda_{\pi'}}T_{v_{\pi'}}(E_\lambda), \quad \pi' = x^{\lambda_{\pi'}}v_{\pi'} \in \Pi', \tag{35}$$

$$E_{s_i(\lambda)} = \left(\widetilde{T}_i + \frac{1-t_i}{1-q^{\langle\lambda,\alpha_i'^\vee\rangle-\langle\alpha_i',w_\lambda(\rho'^\vee)\rangle}}\right)E_\lambda, \quad \langle\lambda,\alpha_i'^\vee\rangle > 0,\ i \neq 0',\ t_i' = 1, \tag{36}$$

$$E_{s_\phi(\lambda)+\phi} = t_\phi' q^{-\langle\phi,w_\lambda(\rho^\vee)\rangle}\left(\widetilde{T}_{0'} + \frac{1-t_{0'}}{1-q^{1-\langle\lambda,\theta'\rangle+\langle\theta',w_\lambda(\rho'^\vee)\rangle}}\right)E_\lambda, \quad \langle\lambda,\phi^\vee\rangle < 1,\ t_{0'}' = 1. \tag{37}$$

*If $t_i' \neq 1$,* (36) *becomes instead*

$$E_{s_i(\lambda)} = \left(\widetilde{T}_i + \frac{1-t_it_i'+(t_i'-t_i)q^{\langle\lambda,\alpha_i'^\vee\rangle-\langle\alpha_i',w_\lambda(\rho'^\vee)\rangle}}{1-q^{2(\langle\lambda,\alpha_i'^\vee\rangle-\langle\alpha_i',w_\lambda(\rho'^\vee)\rangle)}}\right)E_\lambda, \tag{38}$$

*with a corresponding modification to* (37) *if $t_{0'}' \neq 1$.*

The base of the recurrence is $E_\lambda = x^\lambda$ for $\lambda$ minuscule, *i.e.*, $\langle\lambda,\alpha_i^\vee\rangle \geq 0$ for $i \neq 0$ and $\langle\lambda,\phi^\vee\rangle \leq 1$. With this base, (35) is not essential to the recurrence, but it is often useful nevertheless.

To prove Corollary 6.15, first observe that the map $X \to \widetilde{Y}_\mathbb{Q}$, $\lambda \mapsto \Lambda_0^\vee + \lambda$ is equivariant with respect to the action of $W_e'$ on $\widetilde{Y}$ and on $X = W_e'/W_0$. Then Proposition 6.13 and Corollary 6.11 imply that $\Psi_i(E_\lambda)$ (resp. $\pi'(E_\lambda)$) is a scalar multiple of $E_{s_i(\lambda)}$ (resp. $E_{\pi'(\lambda)} = E_{v_{\pi'}(\lambda)+\lambda_{\pi'}}$).

The action of $\Pi'$ on $X = W_e'/W_0$ preserves the Bruhat order. Assuming by induction that (35) holds for $\nu < \lambda$, we conclude that $\pi' = x^{\lambda_{\pi'}}T_{v_{\pi'}}$ carries $\mathbb{Q}(q,t)\{x^\nu :$

$\nu < \lambda\}$ into $\mathbb{Q}(q,t)\{x^\nu : \nu < v_{\pi'}(\lambda) + \lambda_{\pi'}\}$. Hence the coefficient of $x^{v_{\pi'}(\lambda)}$ in $T_{v_{\pi'}}(x^\lambda)$ determines the scalar factor in (35). For $\langle\lambda, \alpha_i^\vee\rangle > 0$ (resp. $\langle\lambda, \phi^\vee\rangle < 1$), we have $s_i w_\lambda > w_\lambda$ (resp. $s_{0'} w_\lambda > w_\lambda$). We may assume by induction that $T_i$ (resp. $T_{0'}$) leaves invariant the space $\mathbb{Q}(q,t)\{x^\nu, x^{s_i(\nu)} : \nu < \lambda\}$. For $i \neq 0'$ and $s_i(\lambda) > \lambda$, we have $[x^{s_i(\lambda)}]\widetilde{T}_i(x^\lambda) = 1$, giving (36), and the coefficient of $x^{s_\phi(\lambda)+\phi}$ in $\widetilde{T}_{0'}(x^\lambda)$ determines the scalar factor in (37). The next lemma supplies the missing scalar factors.

**Lemma 6.16.** (i) *We have* $[x^{v_{\pi'}(\lambda)}]T_{v_{\pi'}}(x^\lambda) = q^{\langle\lambda_{(\pi'^{-1})}, w_\lambda(\rho^\vee)\rangle}$ *for any* $\pi' \in \Pi'$.

(ii) *For* $\langle\lambda, \phi^\vee\rangle < 1$, *we have* $[x^{s_\phi(\lambda)+\phi}]\widetilde{T}_{0'}(x^\lambda) = t_\phi'^{-1} q^{\langle\phi, w_\lambda(\rho^\vee)\rangle}$.

*Proof.* (i) Let $B = (R_0)_+ \cap v_{\pi'}^{-1}(-(R_0)_+)$. We claim that for any $\alpha \in (R_0)_+$, $\langle\lambda_{(\pi'^{-1})}, \alpha^\vee\rangle = 1$ if $\alpha \in B$, 0 otherwise. Then Proposition 5.15 gives

$$[x^{v_{\pi'}(\lambda)}]T_{v_{\pi'}}(x^\lambda) = \prod_{\alpha\in(R_0)_+} u_\alpha^{\langle\lambda_{(\pi'^{-1})},\alpha^\vee\rangle\sigma(-\langle\lambda,\alpha^\vee\rangle)} = q^{\langle\lambda_{(\pi'^{-1})}, w_\lambda(\rho^\vee)\rangle}$$

by the argument in the proof of Proposition 6.9.

As to the claim, if $v_{\pi'} = 1$, then $B = \emptyset$ and $\langle\lambda_{(\pi'^{-1})}, \alpha^\vee\rangle = 0$ for all $\alpha$. Otherwise, $\lambda_{(\pi'^{-1})} = -v_{\pi'}^{-1}(\lambda_{\pi'})$, and $\langle\lambda_{(\pi'^{-1})}, \alpha^\vee\rangle = -\langle\lambda_{\pi'}, v_{\pi'}(\alpha^\vee)\rangle \in \{0,1\}$ for all $\alpha \in (R_0)_+$, since $\lambda_{(\pi'^{-1})}$ is minuscule. Now, $v_{\pi'}(\alpha_j^\vee) = -\phi^\vee$, where $\pi'^{-1}(\alpha_0') = \alpha_j'$, and $v_{\pi'}(\alpha_i^\vee)$ is a simple coroot for $i \neq j$. Since $v_{\pi'} \neq 1$, we have $\langle\lambda_{\pi'}, \phi^\vee\rangle = 1$, and it follows that $\langle\lambda_{(\pi'^{-1})}, \alpha_i^\vee\rangle = \delta_{ij}$. Given $\alpha \in (R_0)_+$, if $\langle\lambda_{(\pi'^{-1})}, \alpha^\vee\rangle = 1$, then $v_{\pi'}(\alpha) \in -(R_0)_+$ since $\lambda_{\pi'} \in X_+$. Conversely, if $\langle\lambda_{(\pi'^{-1})}, \alpha^\vee\rangle = 0$, the coefficient of $\alpha_j^\vee$ in $\alpha^\vee$ must be zero, hence $v_{\pi'}(\alpha) \in (R_0)_+$.

(ii) Let $B = (R_0)_+ \cap s_\phi(-(R_0)_+)$. The operator $s_\phi T_{s_\phi}$ is lower-triangular by Proposition 5.15, hence so is $T_{s_\phi}^{-1} s_\phi$, and $[x^{s_\phi(\lambda)+\phi}]T_{s_\phi}^{-1}(x^{\lambda-\phi})$ is inverse to

$$[x^{\lambda-\phi}]T_{s_\phi}(x^{s_\phi(\lambda)+\phi}) = \prod_{\alpha\in B} u_\alpha^{\sigma(-\langle s_\phi(\lambda)+\phi\,,\,\alpha^\vee\rangle)} = \prod_{\alpha\in B} u_\alpha^{\sigma(\langle\lambda-\phi\,,\,\alpha^\vee\rangle)},$$

using $s_\phi(B) = -B$ in the last equation. Now, $\phi$ is short and dominant, hence $\langle\phi, \alpha^\vee\rangle \in \{0,1\}$ for $\alpha \in (R_0)_+ \setminus \{\phi\}$. Moreover, $s_\phi(\alpha^\vee) = \alpha^\vee - \langle\phi, \alpha^\vee\rangle\phi^\vee$, and since $\phi^\vee$ is the highest coroot, this implies that $\langle\phi, \alpha^\vee\rangle > 0$ if and only if $s_\phi(\alpha) \in -(R_0)_+$. Thus for $\alpha \in (R_0)_+ \setminus \{\phi\}$, we have $\langle\phi, \alpha^\vee\rangle = 1$ if $\alpha \in B$, 0 otherwise. Since $\langle\lambda, \phi^\vee\rangle \leq 0$, it follows that

$$\begin{aligned}[x^{s_\phi(\lambda)+\phi}]\widetilde{T}_{0'}(x^\lambda) &= u_\phi'[x^{s_\phi(\lambda)+\phi}]T_{s_\phi}^{-1}(x^{\lambda-\phi}) = u_\phi' u_\phi \prod_{\alpha\in B\setminus\{\phi\}} u_\alpha^{-\sigma(\langle\lambda,\alpha^\vee\rangle-1)} \\ &= (u_\phi'/u_\phi) \prod_{\alpha\in(R_0)_+} u_\alpha^{\langle\phi,\alpha^\vee\rangle\sigma(-\langle\lambda,\alpha^\vee\rangle)} = t_\phi'^{-1} q^{\langle\phi, w_\lambda(\rho^\vee)\rangle}.\end{aligned}$$

□

**6.17.** Suppose $\widetilde{X}$ is dual to an untwisted type. Then $X, Y$ are of the same type, $\phi = \theta$, $\phi' = \theta'$, $s_\phi = s_{\phi'}$, and in $\mathcal{H}$ we have the identities $T_{0'} = x^\phi T_0^{-1} y^{\phi'} + u_{0'} - u_{0'}^{-1}$ and $\pi' = x^{\lambda_{\pi'}} \pi y^{\lambda_{(\pi^{-1})}}$ for $\pi' \in \Pi'$, $\pi \in \Pi$ such that $v_{\pi'} = v_\pi$. Using these identities, (35) and (38) for $i = 0'$ become

$$E_{v_{\pi'}(\lambda)+\lambda_{\pi'}} = q^{-(\lambda,\lambda_{(\pi^{-1})})} x^{\lambda_{\pi'}} \pi(E_\lambda), \quad \pi' \in \Pi',\ \pi \in \Pi,\ v_\pi = v_{\pi'}$$

$$E_{s_\theta(\lambda)+\theta} = q^{1-(\lambda,\theta')} \left( u_\theta x^{-\alpha_0} T_0^{-1} + \frac{(u_\theta/u_0' - u_\theta u_0') + (u_\theta/u_{0'} - u_\theta u_{0'}) q^r}{1 - q^{2r}} \right) E_\lambda,$$

where $r = 1 - (\lambda, \theta') + \langle \theta', w_\lambda(\rho'^\vee) \rangle$. Note that the second formula simplifies to an analog of (36) if $u_{0'} = u_0'$.

**6.18.** Although our chief concern is with non-symmetric Macdonald polynomials, let us say a little about the symmetric version. Given $\lambda \in X_+$, let $V_\lambda = \mathbb{Q}(q,t)\{E_\nu : \nu \in W_0(\lambda)\}$. By Corollaries 6.14, 6.15, $V_\lambda$ is an $\mathcal{H}(W_0)$-submodule of $\mathbb{Q}(q,t)X$. It follows that there is a unique $W_0$-invariant element $P_\lambda \in V_\lambda$ such that $[x^\lambda]P_\lambda = 1$. The $P_\lambda$ are *symmetric Macdonald polynomials*. They are orthogonal and are joint eigenfunctions of all $W_0$-invariant operators $f(y) \in (\mathbb{Q}(q,t)Y)^{W_0}$. The coefficients of $P_\lambda$ in terms of the $E_\nu$ can be determined explicitly using Corollary 6.15.

The $P_\lambda$ are also orthogonal with respect to *Macdonald's inner product*, which is a symmetrization of $\langle \cdot\,, \cdot \rangle_0$. They were originally defined by Macdonald [14], [15] in terms of this orthogonality. When $t_i = q^{(\alpha_i, \alpha_i')/2}$, they specialize to the irreducible characters of the algebraic group $G$ with weight lattice $X$ and root system $Q_0$. Other specializations yield Hall–Littlewood and Jack polynomials, and spherical functions for classical and $p$-adic symmetric spaces.

For $\mathrm{GL}_n$, the $P_\lambda$ are symmetric polynomials in $x_1, \ldots, x_n$, with coefficients in $\mathbb{Q}(q,t)$. As $n \to \infty$, they converge to symmetric functions $P_\lambda(x; q, t)$ in infinitely many variables $x_i$. A transformed and renormalized variant $\widetilde{H}_\lambda(x; q, t)$ of $P_\lambda(x; q, t)$ was the subject of Macdonald's *positivity conjecture*, proved in [6] by identifying $\widetilde{H}_\lambda(x; q, t)$ with the character of the fiber of a certain vector bundle on the Hilbert scheme $H$ of 0-dimensional subschemes in $\mathbb{C}^2$, at a distinguished point of $H$ corresponding to $\lambda$.

**6.19.** Macdonald polynomials for the maximally non-reduced extensions of affine root systems of type $\widetilde{C}_n^\vee$ are *Koornwinder polynomials*. Their coefficients belong to $\mathbb{Q}(t_0, t_0', t_n, t_n', t_1, q)$. Specializing the five $t$ parameters in various ways yields most Macdonald polynomials for the infinite families of affine root systems.

## 7. A combinatorial formula

**7.1.** From Corollary 6.15 and the definition of the operators $\widetilde{T}_i$ it is clear that for a reduced affine root system, $E_\lambda$ can be expressed as a sum of terms of the form

$$\pm x^\mu q^r t^s \prod_j \frac{1-t_{i_j}}{1-q^{a_j}t^{b_j}},$$

where $t^s$, $t^{b_j}$ stand for monomials in the parameters $t_i$. It may be conjectured, at least for equal parameters $t_i = t$, that $E_\lambda$ is a *positive* sum of such terms. With Haglund and Loehr [5], we proved this for type $\widetilde{A}_{n-1}$ by means of a combinatorial formula, which we will now present (referring the reader to [5] for the proof). Some of the combinatorial structure is the same as in Knop and Sahi's earlier formula [10] for non-symmetric Jack polynomials, but the lift to Macdonald polynomials requires more ingredients.

**7.2.** Take $X = Y = \mathbb{Z}^n$ the root system of $\mathrm{GL}_n$, as in Example 2.5. The pairing $(X, Y) \to \mathbb{Z}$ (§4.4) is the standard inner product on $\mathbb{Z}^n$. We have $\phi = \theta = \phi' = \theta' = e_1 - e_n$, and $\Pi = \Pi'$ is cyclic, with generator $\pi'$ acting on $X = W'_e/W_0$ by

$$\pi'(\lambda) = (\lambda_n + 1, \lambda_1, \ldots, \lambda_{n-1}).$$

To $\pi'$ corresponds an element $\pi \in \Pi$ such that $v_\pi = v_{\pi'}$, which acts on $\mathbb{Q}(q,t)\widetilde{X}$ by

$$\pi(x^\lambda) = q^{-\lambda_n} x^{(\lambda_n, \lambda_1, \ldots, \lambda_{n-1})}, \quad \text{or} \quad \pi f(x_1, \ldots, x_n) = f(x_2, \ldots, x_n, x_1/q).$$

We have $\lambda_{\pi'} = \lambda_\pi = e_1$, $\lambda_{(\pi'^{-1})} = \lambda_{(\pi^{-1})} = -e_n$.

The simple roots are all $W$-conjugate, so there is a single parameter $t_i = t$ for all $i$. For $i \neq 0$, the operators $\widetilde{T}_i$ (§5.13, 6.12) are given by

$$\widetilde{T}_i = t s_i - \frac{1-t}{1-x_i/x_{i+1}}(1-s_i), \tag{39}$$

where $s_i$ is the transposition $x_i \leftrightarrow x_{i+1}$. The analogous formula for $i = 0$ has $qx_n/x_1$ in place of $x_i/x_{i+1}$, and $s_0$ acts as $x_1 \mapsto qx_n$, $x_n \mapsto x_1/q$.

Let $\overline{\lambda}$ be the rearrangement of $(1, 2, \ldots, n)$ such that $\overline{\lambda}_i > \overline{\lambda}_j$ if and only if $\lambda_i > \lambda_j$, for $i < j$. Then $w_\lambda(\rho^\vee) = -k\overline{\lambda}$, modulo a constant vector. From §6.17 and (36), we obtain *Knop's recurrence*, which determines $E_\lambda$ for all $\lambda \in X$:

$$E_{(0,\ldots,0)} = 1, \tag{40}$$

$$E_{(\lambda_n+1, \lambda_1, \ldots, \lambda_{n-1})} = q^{\lambda_n} x_1 E_\lambda(x_2, \ldots, x_n, x_1/q), \tag{41}$$

$$E_{s_i(\lambda)} = \left(\widetilde{T}_i + \frac{1-t}{1-q^{\lambda_i-\lambda_{i+1}}t^{\overline{\lambda}_i-\overline{\lambda}_{i+1}}}\right) E_\lambda, \quad \lambda_i > \lambda_{i+1},\ i \neq 0. \tag{42}$$

**7.3.** By (41), we have $E_{\lambda+(r,r,\dots,r)} = (x_1 \dots x_n)^r E_\lambda$. Without loss of generality, therefore, we restrict attention to *compositions* $\lambda$ such that $\lambda_i \geq 0$ for all $i$. The *column diagram* of $\lambda$ is

$$\mathrm{dg}(\lambda) = \{(i, j) \in \mathbb{N}^2 : 1 \leq i \leq n,\ 1 \leq j \leq \lambda_i\},$$

pictured as an array of $n$ columns of boxes, with $\lambda_i$ boxes in column $i$. The *augmented diagram* of $\lambda$,

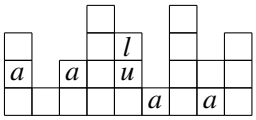
$$\widehat{\mathrm{dg}}(\lambda) = \mathrm{dg}(\lambda) \cup \{(i, 0) : 1 \leq i \leq n\},$$

is formed by adjoining an extra box in row 0 at the bottom of each column. The *arm* and *leg* of a box $u = (i, j) \in \mathrm{dg}(\lambda)$ are defined by

$$l(u) = \mu_i - j,$$
$$a(u) = |\{(i', j) \in \mathrm{dg}(\lambda) : i' < i,\ \lambda_{i'} \leq \lambda_i\} \cup \{(i', j-1) \in \widehat{\mathrm{dg}}(\lambda) : i' > i,\ \lambda_{i'} < \lambda_i\}|.$$

In the example below, the box marked $l$ contributes to the leg of $u$, and those marked $a$ contribute to the arm, giving $l(u) = 1$, $a(u) = 4$.

$\lambda = (2, 0, 1, 3, 2, 0, 3, 1, 2)$, $\widehat{\mathrm{dg}}(\lambda) =$




With these definitions, if $\lambda_i > \lambda_{i+1}$ and $u = (i,\ \mu_{i+1} + 1)$, equation (42) can be written

$$E_{s_i(\lambda)} = \Big(\widetilde{T}_i + \frac{1-t}{1 - q^{l(u)+1} t^{a(u)}}\Big) E_\lambda.$$

A *filling of* $\lambda$ is a function $\sigma : \mathrm{dg}(\lambda) \to [n] = \{1, \dots, n\}$. Its associated *augmented filling* is the extension $\hat{\sigma} : \widehat{\mathrm{dg}}(\lambda) \to [n]$ of $\sigma$ such that $\hat{\sigma}((j, 0)) = j$ for $j = 1, \dots, n$. Distinct boxes $u = (i, j),\ v = (i', j') \in \widehat{\mathrm{dg}}(\lambda)$ *attack* each other if either

(a) they are in the same row, $j = j'$, or

(b) they are in consecutive rows, and the lower box is to the right of the upper box: $j' = j - 1$ and $i < i'$ or vice versa.

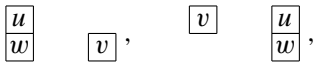
A filling $\sigma$ is *non-attacking* if $\hat{\sigma}(u) \neq \hat{\sigma}(v)$ whenever $u$ and $v$ attack each other (non-attacking fillings are called *admissible* in [10]).

**7.4.** Let $d(u) = (i, j-1)$ denote the box directly below a box $u = (i, j)$. A box $u \in \mathrm{dg}(\lambda)$ is a *descent* of a filling $\sigma$ if $\hat{\sigma}(u) > \hat{\sigma}(d(u))$. Define

$$\mathrm{Des}(\sigma) = \{\text{descents of } \sigma\}, \quad \mathrm{maj}(\sigma) = \sum_{u \in \mathrm{Des}(\sigma)} (l(u) + 1).$$

The *reading order* is the total ordering $<$ of the boxes in $\widehat{\mathrm{dg}}(\lambda)$ row by row, from top to bottom, and from *right to left* within each row. A *triple* consists of three boxes

$u < v < w = d(u)$ in $\widehat{\mathrm{dg}}(\lambda)$, as shown:



Type I Type II

with the proviso that the column containing $u$, $w$ is strictly taller than the column containing $v$ in Type I, and weakly taller in Type II, *i.e.*, $v$ contributes to the arm of $u$. A *co-inversion triple* of $\sigma$ is a triple such that $\sigma(u) < \sigma(v) < \sigma(w)$ or $\sigma(v) < \sigma(w) < \sigma(u)$ or $\sigma(w) < \sigma(u) < \sigma(v)$. Define

$$\mathrm{coinv}(\sigma) = |\{\text{co-inversion triples of } \sigma\}|.$$

**Example 7.5.** The figure below shows the augmentation $\hat{\sigma}$ of a non-attacking filling $\sigma$ of $\lambda = (2, 1, 3, 0, 0, 2)$.

$$\hat{\sigma} = \begin{array}{|c|c|c|c|c|c|}\hline \multicolumn{1}{|c|}{\textcircled{6}} & & \multicolumn{1}{c|}{2} & & & \\ \hline \multicolumn{1}{|c|}{1} & \multicolumn{1}{c|}{2} & \multicolumn{1}{c|}{\textcircled{4}} & & & \multicolumn{1}{c|}{5} \\ \hline \multicolumn{1}{|c|}{1} & \multicolumn{1}{c|}{2} & \multicolumn{1}{c|}{3} & & & \multicolumn{1}{c|}{5} \\ \hline \multicolumn{1}{|c|}{1} & \multicolumn{1}{c|}{2} & \multicolumn{1}{c|}{3} & \multicolumn{1}{c|}{4} & \multicolumn{1}{c|}{5} & \multicolumn{1}{c|}{6} \\ \hline \end{array}.$$

The circled boxes are $\mathrm{Des}(\sigma)$, giving $\mathrm{maj}(\sigma) = 3$. Row 0 is the bottom row. There are two co-inversion triples, one of Type I formed by the 3 and the 5 in row 1 with the 4 in row 2, and one of Type II formed by the 6 and the 4 in row 2 with the 3 in row 1, giving $\mathrm{coinv}(\sigma) = 2$.

**Theorem 7.6.** *The Macdonald polynomials $E_\lambda$ for* $\mathrm{GL}_n$ *are given by*

$$E_\lambda = \sum_{\substack{\sigma:\,\lambda\to[n] \\ \textit{non-attacking}}} x^\sigma q^{\mathrm{maj}(\sigma)} t^{\mathrm{coinv}(\sigma)} \prod_{\substack{u\in\mathrm{dg}(\lambda) \\ \hat{\sigma}(u)\neq\hat{\sigma}(d(u))}} \frac{1-t}{1-q^{l(u)+1}t^{a(u)+1}}, \qquad (43)$$

*where* $x^\sigma = \prod_{u\in\mathrm{dg}(\lambda)} x_{\sigma(u)}$.

**7.7.** Earlier, in [4], we gave a combinatorial formula for the symmetric Macdonald polynomials $P_\lambda$ for $\mathrm{GL}_n$, which had originally been conjectured by Haglund [3]. The combinatorial statistics $\mathrm{coinv}(\sigma)$ and $\mathrm{maj}(\sigma)$ first appeared in the formula for the symmetric case, which is expressed similarly as a sum over fillings of a diagram. Our work in the symmetric case relies heavily on the special theory of $\mathrm{GL}_n$ Macdonald polynomials in the $n \to \infty$ stable limit. It seems likely that the non-symmetric formula will provide better clues as to what we might expect for other root systems.

**7.8.** The proof of Theorem 7.6 is a direct verification that (43) satisfies Knop's recurrence (40)–(42). It is not difficult to check (41), and (40) is trivial. The hard part is to verify (42). In fact, we were only able to do it in the special case that $\lambda_{i+1} = 0$, which fortunately is enough. The difficulty lies in applying the operator $\widetilde{T}_i$ in (39) to (43), which is intractable if attempted head-on. To get around this, we recast (42) as asserting that certain expressions related to (43) are $s_i$-invariant. This is proved with the help of a symmetry lemma which originated in the theory of *LLT polynomials* [11], [12], and was also at the heart of our work in [4]. We invite the reader to consult [5] for more detail.

## References


[1] Cherednik, Ivan, Double affine Hecke algebras, Knizhnik-Zamolodchikov equations, and Macdonald's operators. *Internat. Math. Res. Notices* **(1992)** (9) (1992), 171–180.

[2] —, Nonsymmetric Macdonald polynomials. *Internat. Math. Res. Notices* **(1995)** (10) (1995), 483–515.

[3] Haglund, J., A combinatorial model for the Macdonald polynomials. *Proc. Nat. Acad. Sci. U.S.A.* **101** (46) (2004), 16127–16131.

[4] Haglund, J., Haiman, M., and Loehr, N., A combinatorial formula for Macdonald polynomials. *J. Amer. Math. Soc.* **18** (3) (2005), 735–761.

[5] —, A combinatorial formula for non-symmetric Macdonald polynomials. Preprint, 2006; arXiv:math.CO/0601693.

[6] Haiman, Mark, Hilbert schemes, polygraphs and the Macdonald positivity conjecture. *J. Amer. Math. Soc.* **14** (4) (2001), 941–1006.

[7] Ion, Bogdan, Involutions of double affine Hecke algebras. *Compositio Math.* **139** (1) (2003), 67–84.

[8] Victor G. Kac, *Infinite-dimensional Lie algebras*. Third ed., Cambridge University Press, Cambridge 1990.

[9] Kazhdan, David, and Lusztig, George, Equivariant $K$-theory and representations of Hecke algebras. II. *Invent. Math.* **80** (2) (1985), 209–231.

[10] Knop, Friedrich, and Sahi, Siddhartha, A recursion and a combinatorial formula for Jack polynomials. *Invent. Math.* **128** (1) (1997), 9–22.

[11] Lascoux, Alain, Leclerc, Bernard, and Thibon, Jean-Yves, Ribbon tableaux, Hall-Littlewood functions, quantum affine algebras, and unipotent varieties. *J. Math. Phys.* **38** (2) (1997), 1041–1068.

[12] Leclerc, Bernard, and Thibon, Jean-Yves, Littlewood-Richardson coefficients and Kazhdan-Lusztig polynomials.In *Combinatorial methods in representation theory* (Kyoto, 1998), Adv. Stud. Pure Math. 28, Kinokuniya, Tokyo 2000, 155–220.

[13] Macdonald, I. G., Some conjectures for root systems. *SIAM J. Math. Anal.* **13** (6) (1982), 988–1007.

[14] —, *A new class of symmetric functions*. In *Actes du 20e Séminaire Lotharingien*, vol. 372/S-20, Publications I.R.M.A., Strasbourg 1988, 131–171.

[15] —, Orthogonal polynomials associated with root systems. In *Sém. Lothar. Combin.* **45** (2000/01), Art. B45a, 40 pp. (electronic).

[16] —, *Affine Hecke algebras and orthogonal polynomials*. Cambridge Tracts in Math. 157, Cambridge University Press, Cambridge 2003.

[17] Opdam, Eric M., Harmonic analysis for certain representations of graded Hecke algebras. *Acta Math.* **175** (1) (1995), 75–121.



Department of Mathematics, University of California, Berkeley, CA, U.S.A.

E-mail: mhaiman@math.berkeley.edu

# Poisson cloning model for random graphs

Jeong Han Kim

**Abstract.** In the random graph $G(n, p)$ with $pn$ bounded, the degrees of the vertices are almost i.i.d. Poisson random variables with mean $\lambda := p(n-1)$. Motivated by this fact, we introduce the Poisson cloning model $G_{\mathrm{PC}}(n, p)$ for random graphs in which the degrees are i.i.d. Poisson random variables with mean $\lambda$.

We first establish a theorem that shows that the new model is equivalent to the classical model $G(n, p)$ in an asymptotic sense. Next, we introduce a useful algorithm to generate the random graph $G_{\mathrm{PC}}(n, p)$, called the cut-off line algorithm. Then $G_{\mathrm{PC}}(n, p)$ equipped with the cut-off line algorithm enables us to very precisely analyze the sizes of the largest component and the $t$-core of $G(n, p)$. This new approach for the problems yields not only elegant proofs but also improved bounds that are essentially best possible.

We also consider the Poisson cloning model for random hypergraphs and the $t$-core problem for random hypergraphs.

**Mathematics Subject Classification (2000).** Primary 05C80; Secondary 05D40.

**Keywords.** Random graph, giant component, core, Poisson distribution.

## 1. Introduction

The notion of a random graph was first introduced in 1947 by Erdős [14] to show the existence of a graph with a certain Ramsey property. A decade later, the theory of random graphs began with the paper entitled *On Random Graphs I* by Erdős and Rényi [15], and they further developed the theory in a series of papers [16], [17], [18], [19], [20]. Since then, the subject has become one of the most active research areas. Many researchers have devoted themselves to studying various properties of random graphs, such as the emergence of the giant component [16], [5], [32], the connectivity [15], [17], [11], the existence of perfect matching [18], [19], [20], [11], the existence of Hamiltonian cycle(s) [31], [6], [10], the $k$-core problem [6], [34], [38], and the graph invariants like the independence number [9], [36] and the chromatic number [39], [8], [33]. (The list of references here is far from being exhaustive.)

There are two canonical models for random graphs, both of which were originated in the simple model introduced in [14]. In the binomial model $G(n, p)$ on a set $V$ of $n$ vertices, each of $\binom{n}{2}$ possible edges is in the graph with probability $p$, independently of other edges. Thus, the probability of $G(n, p)$ being a fixed graph $G$ with $m$ edges is $p^m(1-p)^{\binom{n}{2}-m}$. The uniform model $G(n, m)$ on $V$ is a graph chosen uniformly at random from the set of all graphs on $V$ with $m$ edges. Hence $G(n, m)$ becomes a fixed

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

graph $G$ with probability $\binom{\binom{n}{2}}{m}^{-1}$, provided $G$ has $m$ edges. Most of the asymptotic behavior of the two models is almost identical if their expected numbers of edges are the same. (See Proposition 1.13 in [27].) The random graph process, in which random edges are added one by one, is also extensively studied. For more about models and/or basics of random graphs we recommend two books with the same title: *Random Graphs* by Bollobás [7] and by Janson, Łuczak and Ruciński [27].

The phase transition phenomenon is one of the interesting topics in random graphs. Specifically, the phase transition phenomena regarding the emergences of the giant (connected) component and the $t$-core have attracted much attention. In the monumental paper *On the Evolution of Random Graphs* [16], Erdős and Rényi proved that, for the size $\ell_1(n, p)$ of the largest component of $G(n, p)$,

$$\ell_1(n, p) = \begin{cases} O(\log n), & \text{if } \limsup_{n\to\infty} p(n-1) < 1, \\ (1 + o(1))\theta_\lambda n, & \text{if } \lim_{n\to\infty} pn = \lambda > 1, \end{cases}$$

where $\theta_\lambda$ is the positive solution of the equation $1 - \theta - e^{-\lambda\theta} = 0$ and all other components are of size $O(\log n)$.

Why does the size of the largest component change so dramatically around $\lambda = 1$? It was Karp [28] who nicely explained the reason. To find a component $C(v)$ of a fixed vertex $v$ in $G(n, p)$, one may first expose the vertices that are adjacent to $v$ and keep on repeating the same procedure by taking each of those adjacent vertices. Initially, $v$ is active and all other vertices are neutral. At each step we take an active vertex $w$ and expose all neutral vertices adjacent to $w$. This can be done by checking $\{w, w'\} \in G(n, p)$ or not for all neutral vertices $w'$. Then activate all neutral vertices that are adjacent to $w$. The vertex $w$ is no longer active, and only non-activated neutral vertices remain neutral. The process terminates when there is no more active vertex left. Clearly, the process will stop after finding all the vertices in the component containing $v$. If the number of neutral vertices does not decrease so fast, the number of newly activated vertices is close to the binomial random variable $\mathrm{Bin}(n-1, p)$, where

$$\Pr[\mathrm{Bin}(n-1, p) = \ell] = \binom{n-1}{\ell} p^\ell (1-p)^{n-1-\ell}.$$

Particularly, the mean number is close to $pn$. If $pn \le 1 - \delta$ for a fixed $\delta > 0$, then the process is expected to die out quickly almost every time. Thus, all $C(v)$ are expected to be small. If $pn \ge 1 + \delta$, then the process may survive forever with positive probability. Hence, $C(v)$ can be large with positive probability. As there are $n$ trials, at least one of the $C(v)$'s is expected to be large. Applying this approach to the random directed graph, Karp was able to prove a phase transition phenomenon for the size of the largest strong component.

Notice that, when $pn = \Theta(1)$, the distribution of $\mathrm{Bin}(n-1, p)$ is very close the Poisson distribution with parameter $\lambda := p(n-1)$. Hence we may further expect that the process described above could be approximated by the Galton–Watson branching

process defined by a Poisson random variable $\mathrm{Poi}(\lambda)$ with mean $\lambda$, where

$$\Pr[\mathrm{Poi}(\lambda) = \ell] = e^{-\lambda}\frac{\lambda^{\ell}}{\ell!}.$$

Generally, the Galton–Watson branching process defined by a random variable $X$ starts with a single unisexual organism. The organism will give birth to $X_1$ children, where $X_1$ is a random variable with the same distribution as $X$. The same but independent birth process continues from each of the children, the grandchildren and so on, until no more descendant exists. (For more information regarding Galton–Watson branching processes, one may refer [4].) For simplicity we call the Galton–Watson branching process defined by $\mathrm{Poi}(\lambda)$ the *Poisson*$(\lambda)$ *branching process*.

**The Poisson cloning model.** To convert the above observation to a rigorous proof, it is needed to overcome or bypass two main obstacles. Firstly, the degrees of vertices of $G(n, p)$ are not exactly i.i.d. Poisson random variables. Though they have the same distribution as $\mathrm{Bin}(n-1, p)$, they are not mutually independent. For example, the sum of all degrees must be even as it is twice the number of edges, which cannot be guaranteed if the degrees are independent. Secondly, the number of neutral vertices keeps decreasing. Even if both obstacles do not cause substantial differences in many cases, one needs at least to keep tracking small differences for rigorous proofs. Since this kind of small differences occurs almost everywhere in the analysis, it sometimes makes rigorous analysis significantly difficult, if not impossible.

As an approach to bypass the first obstacle, we introduce the Poisson cloning model $G_{\mathrm{PC}}(n, p)$ for random graphs in which the degrees are i.i.d. Poisson random variables with mean $\lambda = p(n-1)$. Moreover, the new model is equivalent to the classical model $G(n, p)$ in an asymptotic sense. Actually, defining the model is not extremely difficult: First take i.i.d. Poisson $\lambda$ random variables $d(v)'s$ indexed by all vertices in the vertex set $V$. Then take $d(v)$ copies, or clones, of each vertex $v$. If the sum of $d(v)$'s is even, then we generate a uniform random perfect matching on the set of all clones. An edge $\{v, w\}$ is in the random graph $G_{\mathrm{PC}}(n, p)$ if a clone of $v$ is matched to a clone of $w$ in the random perfect matching. If the sum is odd, one may just take a graph with a self loop. Hence the graph is not simple if the sum is odd.

It is also possible to extend the model to uniform hypergraphs, where a $k$-uniform hypergraph on the vertex set $V$ is a collection of subsets of $V$ with size $k$. A graph is then a 2-uniform hypergraph. In the binomial model $H(n, p; k)$ for random $k$-uniform hypergraphs each of $\binom{n}{k}$ edges is in the hypergraph with probability $p$, independently of other edges. The Poisson cloning model for random $k$-uniform hypergraphs is denoted by $H_{\mathrm{PC}}(n, p; k)$. In the next section the Poisson cloning model is defined in detail.

The following theorem shows that the Poisson cloning model is essentially equivalent to the binomial model.

**Theorem 1.1.** *Suppose $k \geq 2$ and $p = \Theta(n^{1-k})$. Then for any collection $\mathcal{H}$ of $k$-uniform simple hypergraphs,*

$$c_1 \Pr[H_{\mathrm{PC}}(n, p; k) \in \mathcal{H}] \leq \Pr[H(n, p; k) \in \mathcal{H}] \leq c_2\big(\Pr[H_{\mathrm{PC}}(n, p; k) \in \mathcal{H}]^{\frac{1}{k}} + e^{-n}\big),$$

*where*

$$c_1 = k^{1/2} e^{\frac{p}{n}\binom{k}{2}\binom{n}{k} + \frac{p^2}{2}\binom{n}{k}} + O(n^{-1/2}), \quad c_2 = \Big(\frac{k}{k-1}\Big)\big(c_1(k-1)\big)^{1/k} + o(1),$$

*and $o(1)$ goes to 0 as $n$ goes to infinity.*

To overcome the second obstacle we present an algorithm, called the cut-off line algorithm, that enables us to generate the Poisson cloning model and analyze problems simultaneously. As a consequence the size of the largest component of $G_{\mathrm{PC}}(n, p)$ can be described very precisely. It is also possible to analyze the size of $t$-core of the random hypergraph $H_{\mathrm{PC}}(n, p; k)$, where the $t$-core of a hypergraph $H$ is the largest subhypergraph of $H$ with minimum degree at least $t$.

**The emergence of the giant component.** After the phase transition result of Erdős and Rényi it remained to determine the size of the largest component when $pn \to 1$. Though Erdős and Rényi suggested that the size $\ell_1(n, p)$ of the largest component could be one of $O(\log n)$, $\Theta(n^{2/3})$, and $\Theta(n)$, Bollobás [5] showed that $\ell_1(n, p)$ increases rather continuously by estimating it quite accurately for $pn - 1 \geq n^{-1/3}\sqrt{\log n}/2$. Later Łuczak [32] was able to estimate $\ell_1(n, p)$ for $pn - 1 \gg n^{-1/3}$.

Before stating the result of Łuczak a convention is introduced: when the expression $x \gg y$ is used as part of the hypotheses, it means 'there exists a (large) constant $K > 0$ so that, if $x \geq Ky \ldots$'. We also denote $\lambda(n, p) = p(n-1)$.

**Theorem 1.2** (Supercritical phase). *Suppose $\lambda = \lambda(p, n) = 1 + \varepsilon$ with $\varepsilon \gg n^{-1/3}$. Then for large enough $n$, with probability at least $1 - 7(\varepsilon^3 n/8)^{-1/9}$*

$$|\ell_1(n, p) - \theta_\lambda n| \leq \frac{n^{2/3}}{5},$$

*and all other components are smaller than $n^{2/3}$.*

Using estimations for the numbers of connected graphs with certain number of vertices and edges and the first and second moment methods, one may also obtain the following result for the subcritical phase.

**Theorem 1.3** (Subcritical phase). *Let $\lambda(n, p) = 1 - \varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$. Then for any positive constant $\delta \leq 1/3$ and large enough $n$, with probability at least $1 - \big(\frac{8}{\varepsilon^3 n}\big)^{\delta/4}$*

$$\left|\ell_1(n, p) - \frac{2\log(\varepsilon^3 n)}{\varepsilon^2}\right| \leq \frac{\delta \log(\varepsilon^3 n)}{\varepsilon^2}.$$

For results regarding the structure of the largest component readers are referred to [27], [32], [25], [35] and references therein.

For Poisson branching processes a duality principle has been known. A pair $(\mu, \lambda)$ with $\mu < 1 < \lambda$ is called a *conjugate pair* if $\mu e^{-\mu} = \lambda e^{-\lambda}$. It is easy to see that $\mu = (1 - \theta_\lambda)\lambda$ for a conjugate pair $(\mu, \lambda)$. For a conjugate pair $(\mu, \lambda)$ the distribution of the Poisson($\lambda$) branching process conditioned that the process dies out is exactly the same as that of the Poisson($\mu$) branching process. (See e.g. [2], p. 164.) A similar duality was observed for the random graph $G(n, p)$ and $G(n^*, p)$ with $\lambda = \lambda(n, p) > 1$ and $n^* = (1 - \theta_\lambda)n$. (Recall that $\lambda(n, p) = p(n - 1)$.) Notice that $1 - \theta_\lambda$ is the extinction probability for the Poisson($\lambda$) branching process. It is known that the component sizes of $G(n^*, p)$ and those of $G(n, p)$ excluding the largest component are the same in an asymptotic sense (see [2]).

The Poisson cloning model $G_{\mathrm{PC}}(n, p)$ equipped with the cut-off line algorithm enables us to not only estimate $\ell_1(n, p)$ more accurately but also to establish a discrete duality principle: in the supercritical phase $\lambda := \lambda(n, p) = 1 + \varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$; $G_{\mathrm{PC}}(n, p)$ can be decomposed by three vertex disjoint graphs $C$, $S$ and $G$ whp (with high probability), where $C$ is a connected graph of size about $\theta_\lambda n$, $S$ is a smaller graph of size about $\varepsilon^{-2} \ll \theta_\lambda n$, and $G$ has the same distribution as $G_{\mathrm{PC}}(n^*, p^*)$ with $n^* \approx (1 - \theta_\lambda)n$ and $p^* \approx p$, which yields $\lambda(n^*, p^*) \approx \mu := (1 - \theta_\lambda)\lambda$. In the subcritical phase $\lambda = 1 - \varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$ the largest component is of size

$$\frac{\log(\varepsilon^3 n) - 2.5 \log\log(\varepsilon^3 n) + O(1)}{-(\varepsilon + \log(1 - \varepsilon))}$$

whp. The precise statements are as follows. We concentrate on the cases $\varepsilon \ll 1$ for which more careful analysis is required. It is believed that the proofs are easily modified for the cases of positive constants $\varepsilon$.

**Theorem 1.4.** *Supercritical phase: Let $\lambda := \lambda(n, p) = 1 + \varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$, $\mu := (1-\theta_\lambda)\lambda$ and $1 \ll \alpha \ll (\varepsilon^3 n)^{1/2}$. Then with probability $1 - e^{-\Omega(\alpha^2)}$ $G_{\mathrm{PC}}(n, p)$ may be decomposed by three vertex disjoint graphs $C$, $S$ and $G$, where $C$ is connected and*

$$\theta_\lambda n - \alpha(n/\varepsilon)^{1/2} \le |C| \le \theta_\lambda n + \alpha(n/\varepsilon)^{1/2},$$

*$|S| \le \frac{\alpha^2}{\varepsilon^2}$, $G$ has the same distribution as $G_{\mathrm{PC}}(n^*, p^*)$ for some $n^*$ and $p^*$ satisfying*

$$(1 - \theta_\lambda)n - \alpha(n/\varepsilon)^{1/2} \le n^* \le (1 - \theta_\lambda)n + \alpha(n/\varepsilon)^{1/2},$$

*and*

$$\mu - \alpha(\varepsilon n)^{-1/2} \le \lambda(n^*, p^*) \le \mu + \alpha(\varepsilon n)^{-1/2}.$$

*Subcritical phase: Suppose $\lambda := \lambda(n, p) = 1 - \varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$. Then the size $\ell_1^{\mathrm{PC}}(n, p)$ of the largest component of $G_{\mathrm{PC}}(n, p)$ satisfies*

$$\Pr\left[\ell_1^{\mathrm{PC}}(n, p) \ge \frac{\log(\varepsilon^3 n) - 2.5 \log\log(\varepsilon^3 n) + c}{-(\varepsilon + \log(1 - \varepsilon))}\right] \le 2e^{-\Omega(c)},$$

*and*

$$\Pr\left[\ell_1^{\mathrm{PC}}(n,p) \le \frac{\log(\varepsilon^3 n) - 2.5\log\log(\varepsilon^3 n) - c}{-(\varepsilon + \log(1-\varepsilon))}\right] \le 2e^{-e^{\Omega(c)}}$$

*for any constant $c > 0$.*

*Inside window: Suppose $\lambda := \lambda(n,p) = 1+\varepsilon$ with $|\varepsilon| = O(n^{1/3})$. Then whp*

$$\ell_1^{\mathrm{PC}}(n,p) = \Theta(n^{2/3}).$$

(*All constants in the $\Omega(\cdot)$'s do not depend on any of $\varepsilon$, $\alpha$ and $c$.*)

By Theorem 1.1 a corollary regarding $G(n,p)$ follows.

**Corollary 1.5.** *Supercritical region: Suppose $\lambda = \lambda(n,p) = 1+\varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$, and $1 \ll \alpha \ll (\varepsilon^3 n)^{1/2}$. Then, in $G(n,p)$,*

$$\Pr[\,|\ell_1(n,p) - \theta_\lambda n| \ge \alpha(n/\varepsilon)^{1/2}] \le 2e^{-\Omega(\alpha^2)}.$$

*Moreover, for the size $\ell_2(n,p)$ of the second largest component and $\varepsilon^* = 1-(1-\theta_\lambda)\lambda$,*

$$\Pr\left[\ell_2(n,p) \ge \frac{\log((\varepsilon^*)^3 n) - 2.5\log\log((\varepsilon^*)^3 n) + c}{-(\varepsilon^* + \log(1-\varepsilon^*))}\right] \le 2e^{-\Omega(c)}$$

*and*

$$\Pr\left[\ell_2(n,p) \le \frac{\log((\varepsilon^*)^3 n) - 2.5\log\log((\varepsilon^*)^3 n) - c}{-(\varepsilon^* + \log(1-\varepsilon^*))}\right] \le 2e^{-e^{\Omega(c)}},$$

*for any constant $c > 0$.*

*Subcritical region: Suppose $\lambda = 1-\varepsilon$ with $n^{-1/3} \ll \varepsilon \ll 1$. Then for any constant $c > 0$,*

$$\Pr\left[\ell_1(n,p) \ge \frac{\log(\varepsilon^3 n) - 2.5\log\log(\varepsilon^3 n) + c}{-(\varepsilon + \log(1-\varepsilon))}\right] \le 2e^{-\Omega(c)}$$

*and*

$$\Pr\left[\ell_1(n,p) \le \frac{\log(\varepsilon^3 n) - 2.5\log\log(\varepsilon^3 n) - c}{-(\varepsilon + \log(1-\varepsilon))}\right] \le 2e^{-e^{\Omega(c)}}.$$

*Inside window: Suppose $\lambda := \lambda(n,p) = 1+\varepsilon$ with $|\varepsilon| = O(n^{1/3})$. Then whp*

$$\ell_1(n,p) = \Theta(n^{2/3}).$$

**The emergence of the $t$-core.** There are at least two possible directions to extend the problem of (connected) component. Observing that the minimum degree in a component must be larger than or equal to 1, one may consider subgraphs with minimum degree at least $t \ge 2$. For a graph $G$, the $t$-core is the largest subgraph with minimum degree at least $t$. Since the minimum degree of the union of two subgraphs is larger than or equal to the smaller minimum degree of the two, the $t$-core of a graph

is unique. It is also easy to see that the $t$-core must be an induced subgraph. For this reason the $t$-core of $G$ sometimes refers to its vertex set. Denote by $V_t(G)$ (the vertex set of) the $t$-core of $G$. As the 1-core $V_1(G)$ is the set of all non-isolated vertices, we consider the cases $t \geq 2$ throughout this paper. If there is no subgraph with minimum degree $t$, the $t$-core is defined to be empty.

Another direction is to consider the $t$-connectivity, where a graph is $t$-connected if the graph remains connected after any $t-1$ vertices are removed. Higher orders of connectivity have been used to understand various structures of graphs. Clearly, if a non-empty subgraph is $t$-connected, then its minimum degree must be $t$ or larger.

In 1984, Bollobás [6] initiated the study of $t$-core, $t \geq 2$, and observed that, provided $t \geq 3$ and $pn$ is larger than a fixed constant, the $t$-core of $G(n, p)$ is non-empty and $t$-connected whp. Łuczak [34] proved that for $t \geq 3$ there is an absolute constant $c$ such that the $t$-core of $G(n, p)$ is either empty, or larger than $cn$ and $t$-connected, whp. In particular, as far as the random graph $G(n.p)$ is concerned, the $t$-core problem is the same as the $t$-connectivity problem. Moreover, if $\lambda(n, p)$ is less than 1, then the $t$-core of $G(n, p)$ is empty whp since the size of the largest component is $O(n^{2/3})$ whp. As $p$ increases while $n$ is fixed, the probability of the $t$-core of $G(n, p)$ being non-empty keeps increasing. Let $p_t(n, \delta)$ be the infimum of all $p$ that makes the probability larger than or equal to a constant $\delta$ with $0 < \delta < 1$. Then Bollobás's result implies that $np_t(n, \delta)$ is bounded from above by a constant. Though $np_t(n, \delta)$ may still have no limit value as $n$ goes to infinity, it seems to be more natural to expect that the limit exists. Furthermore, as it happens often in phase transition phenomena, the limit, if it exists, is also expected to be independent of $\delta$. In other words, the phase transition is expected to be sharp.

For $t = 2$, the 2-core of a graph $G$ is non-empty if and only if $G$ contains a cycle. It is easy to see by the first moment method that $G(n, p)$ with $p = o(1/n)$ does not contain a cycle whp. For a constant $c$ with $0 < c < 1$, $G(n, p)$ may or may not have a cycle with positive probability. In particular, the phase transition for the existence of a non-empty 2-core is not sharp. In the graph process $(G(n, m))_{m=0,1,\ldots}$ in which a random edge is added one by one without repetition, Janson [24] found the limiting distribution for the length of the first cycle, especially he showed that the length is bounded whp. However, the expectation of the length is known to be $\Theta(n^{1/6})$ due to Flajolet et al. [23]. The two facts do not contradict each other, since there are random variables $X$ that are bounded whp, but $E[X]$ is not. For example, $\Pr[X = 1] = 1 - 1/n$ and $\Pr[X = n^2] = 1/n$.

Bollobás [6] proved that, if $t \geq 5$ and $\lambda(n, p) := p(n-1) \geq \max\{67, 2t + 6\}$, then $G(n, p)$ has a non-empty $t$-core. Chvátal [12] introduced the notion of critical $\lambda_t$, without proving existence, satisfying the following. As $n$ goes to infinity,

$$\Pr[G(n, p) \text{ has a non-empty } t\text{-core }] \longrightarrow \begin{cases} 0 & \text{if } \lambda(n, p) < \lambda_t - \delta, \\ 1 & \text{if } \lambda(n, p) > \lambda_t + \delta \end{cases}$$

for any constant $\delta > 0$. He also proved that $\lambda_3 \geq 2.88$ if it exists, and claimed that

$\lambda_4 \geq 4.52$ and $\lambda_5 \geq 6.06$ etc. could be proven by the same method. Pittel, Spencer and Wormald [38] proved a general theorem which implies that $\lambda_t$ exists for fixed $t \geq 3$ and identified their values. We present a slightly weaker version of this result.

For a Poisson random variable $\mathrm{Poi}(\rho)$ with mean $\rho$ let $P(\rho, i) = \Pr[\mathrm{Poi}(\rho) = i]$ and $Q(\rho, i) = \Pr[\mathrm{Poi}(\rho) \geq i]$, i.e.,

$$P(\rho, i) = e^{-\rho}\frac{\rho^i}{i!} \quad \text{and} \quad Q(\rho, i) := \sum_{j=i}^{\infty} P(\rho, j) = e^{-\rho}\sum_{j=i}^{\infty}\frac{\rho^j}{j!},$$

and let

$$\lambda_t = \min_{\rho>0}\frac{\rho}{Q(\rho, t-1)}.$$

**Theorem 1.6.** *Let $t \geq 3$, $\lambda(n, p) = p(n-1)$. Then*

$$\Pr[G(n, p) \text{ has a non-empty } t\text{-core}] \longrightarrow \begin{cases} 0 & \text{if } \lambda(n, p) < \lambda_t - n^{-\delta}, \\ 1 & \text{if } \lambda(n, p) > \lambda_t + n^{-\delta} \end{cases}$$

*for any $\delta \in (0, 1/2)$, and the $t$-core when $\lambda(n, p) > \lambda_t + n^{-\delta}$ has $(1+o(1))Q(\theta_\lambda\lambda, t)n$ vertices, whp, where $\theta_\lambda$ is the largest solution for the equation*

$$\theta - Q(\theta\lambda, t-1) = 0.$$

There are many studies about the $t$-cores of various types of random graphs and random hypergraphs. Fernholz and Ramachandran [21], [22] studied random graph conditions on given degree sequences. Cooper [13] found the critical values for $t$-cores of a uniform multihypergraph with given degree sequences that includes the random $k$-uniform hypergraph $H(n, p\,; k)$. Molloy [37] considered cores for random hypergraphs and random satisfiability problems for Boolean formulas. Recently, S. Janson and M. J. Luczak [26] also gave seemingly simpler proofs for $t$-core problems that contain the result of Pittel, Spencer and Wormald. For more information and techniques used in the above mentioned papers readers are referred to [26].

Using the Poisson cloning model for random hypergraphs together with the cut-off line algorithm we are able to completely analyze the $t$-core problem for the random uniform hypergraph. We also believe that the cut-off line algorithm can be used to analyze the $t$-core problem for random hypergraphs conditioned on certain degree sequences as in [13], [21], [22], [26].

As the 2-core of $G(n, p)$ behaves quite differently from the other $t$-cores of $H(n, p\,; k)$, we exclude the case $k = t = 2$, which will be studied in a subsequent paper. The critical value for the problem turns out to be the minimum $\lambda$ such that there is a positive solution for the equation

$$\theta - Q(\theta^{k-1}\lambda, t-1) = 0. \tag{1.1}$$

It is not difficult to check that the minimum is

$$\lambda_{\mathrm{crt}}(k, t) := \min_{\rho>0}\frac{\rho}{Q(\rho, t-1)^{k-1}}. \tag{1.2}$$

For $\lambda > \lambda_{\mathrm{crt}}(k,t)$, let $\theta_\lambda$ be the largest solution of the equation $\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1) = 0$.

**Theorem 1.7.** *Let $k, t \geq 2$, excluding $k = t = 2$, and $\sigma \gg n^{-1/2}$.*

*Subcritical phase: If $\lambda(n, p; k) := p\binom{n-1}{k-1} = \lambda_{\mathrm{crt}} - \sigma$ is uniformly bounded from below by 0 and $i_0(k,t)$ is the minimum $i$ such that $\binom{i}{k} \geq ti/k$, then*

$$\Pr[V_t(H(n, p; k)) \neq \emptyset] = e^{-\Omega(\sigma^2 n)} + O\big(n^{-(t-1-t/k)i_0(k,t)}\big),$$

*and for any $\delta > 0$,*

$$\Pr[|V_t(H(n, p; k))| \geq \delta n] = e^{-\Omega(\sigma^2 n)} + e^{-\Omega(\delta^{2k/(k-1))} n)}. \tag{1.3}$$

*Supercritical phase: If $\lambda = \lambda(n, p; k) = \lambda_{\mathrm{crt}} + \sigma$ is uniformly bounded from above, then for all $\alpha$ in the range $1 \ll \alpha \ll \sigma n^{1/2}$,*

$$\Pr\big[\big||V_t(n, p; k)| - Q(\theta_\lambda\lambda, t)n\big| \geq \alpha(n/\sigma)^{1/2}\big] = e^{-\Omega(\alpha^2)}, \tag{1.4}$$

*and, for any $i \geq t$ and the sets $V_t(i)$ (resp. $W_t(i)$) of vertices of degree $i$ (resp. larger than or equal to $i$) in the $t$-core,*

$$\Pr\big[\big||V_t(i)| - P(\theta_\lambda\lambda, i)n\big| \geq \delta n\big] \leq 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})},$$

*and*

$$\Pr\big[\big||W_t(i)| - Q(\theta_\lambda\lambda, i)n\big| \geq \delta n\big] \leq 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}.$$

*In particular, for $\lambda = \lambda_{\mathrm{crt}} + \sigma$ and $\rho_{\mathrm{crt}} := \theta_{\lambda_{\mathrm{crt}}(k,t)}\lambda_{\mathrm{crt}}(k,t)$,*

$$|V_t(i)| = \big(1 + O(\sigma^{1/2})\big)^i P(\rho_{\mathrm{crt}}, i)n + O\big((n/\sigma)^{1/2}\log n\big),$$

*with probability $1 - 2e^{-\Omega(\min\{\log^2 n, \sigma^2 n\})}$.*

As one might guess, we will prove a stronger theorem (Theorem 6.2) for the Poisson cloning model $H_{\mathrm{PC}}(n, p; k)$, from which Theorem 1.7 easily follows.

In the next section the Poisson cloning model is defined in detail. The cut-off line algorithm and the cut-off line lemma are presented in Section 3. In Section 4 we study Chernoff type large derivation inequalities that will be used in most of our proofs. In Section 5, a generalized core is defined and the main lemma is presented. Section 6 is devoted to the proof of Theorem 1.7. As the proof of Theorem 1.4 is more sophisticated, we only give the proof ideas in Section 7. We conclude this paper with final remarks in Section 8. Due to the space limitation, many proofs are omitted. They can be found on the author's web site.

## 2. The Poisson cloning model

To construct the Poisson cloning model $G_{\mathrm{PC}}(n, p)$ for random graphs, let $V$ be a set of $n$ vertices. We take i.i.d. Poisson $\lambda = p(n-1)$ random variables $d(v)$, $v \in V$, and then take $d(v)$ copies of each vertex $v \in V$. The copies of $v$ are called *clones of* $v$, or simply *$v$-clones*. Since the sum of Poisson random variables is also a Poisson random variable, the total number $N_\lambda := \sum_{v \in V} d(v)$ of clones is a Poisson $\lambda n$ random variable. It is sometimes convenient to take a reverse, but equivalent, construction. We first take a Poisson $\lambda n = 2p\binom{n}{2}$ random variables $N_\lambda$ and then take $N_\lambda$ unlabelled clones. Each clone is independently labelled as $v$-clone uniformly at random, in the sense that $v$ is chosen uniformly at random from $V$. It is well known that the numbers $d(v)$ of $v$-clones are i.i.d. Poisson random variables with mean $\lambda$.

If $N_\lambda$ is even, the multigraph $G_{\mathrm{PC}}(n, p)$ is defined by generating a (uniform) random perfect matching of those $N_\lambda$ clones, and contracting clones of the same vertex. That is, if a $v$-clone and a $w$-clone are matched, then the edge $\{v, w\}$ is in $G_{\mathrm{PC}}(n, p)$ with multiplicity. In the case that $v = w$, it produces a loop that contributes 2 to the degree of $v$. If $N_\lambda$ is odd, we may define $G_{\mathrm{PC}}(n, p)$ to be any graph with a special loop that, unlike other loops, contributes only 1 to the degree of the corresponding vertex. In particular, if $N_\lambda$ is odd, $G_{\mathrm{PC}}(n, p)$ is not a simple graph.

Strictly speaking, $G_{\mathrm{PC}}(n, p)$ varies depending on how to define it when $N_\lambda$ is odd. However, if only simple graphs are concerned, the case of $N_\lambda$ being odd would not matter. For example, the probability that $G_{\mathrm{PC}}(n, p)$ is a simple graph with a component larger than $0.1n$ does not depend on how $G_{\mathrm{PC}}(n, p)$ is defined when $N_\lambda$ is odd, as it is not a simple graph anyway. Generally, for any collection $\mathcal{G}$ of simple graphs, the probability that $G_{\mathrm{PC}}(n, p)$ is in $\mathcal{G}$ is totally independent of how $G_{\mathrm{PC}}(n, p)$ is defined when $N_\lambda$ is odd. Notice that properties of simple graphs are actually mean collections of simple graphs. Therefore, when properties of simple graphs are concerned, it is not necessary to describe $G_{\mathrm{PC}}(n, p)$ for odd $N_\lambda$.

Here are two specific ways to generate the uniform random matching.

**Example 2.1.** One may keep matching two clones chosen uniformly at random among all unmatched clones.

**Example 2.2.** One may keep choosing his or her favorite unmatched clone, and matching it to a clone selected uniformly at random from all other unmatched clones.

If $N_\lambda$ is even both examples would yield uniform random perfect matchings. If $N_\lambda$ is odd, then each of them would yield a matching and an unmatched clone. We may create the special loop consisting of the vertex for which the unmatched clone is labelled. More specific ways to choose random clones will be described in the next section.

Generally for $k \geq 3$, the Poisson cloning model $H_{\mathrm{PC}}(n, p\,; k)$ for $k$-uniform hypergraphs may be defined in the same way: We take i.i.d. Poisson $\lambda = p\binom{n-1}{k-1}$ random variables $d(v)$, $v \in V$, and then take $d(v)$ clones of each $v$. If $N_\lambda := \sum_{v \in V} d(v)$ is

divisible by $k$, the multihypergraph $H_{\mathrm{PC}}(n, p; k)$ is defined by generating a uniform random perfect matching consisting of $k$-tuples of those $N_\lambda$ clones, and contracting clones of the same vertex. That is, if $v_1$-clone, $v_2$-clone, …, $v_k$-clone are matched in the perfect matching, then the edge $\{v_1, v_2, \dots, v_k\}$ is in $H_{\mathrm{PC}}(n, p\,; k)$ with multiplicity. If $N_\lambda$ is not divisible by $k$, $H_{\mathrm{PC}}(n, p\,; k)$ may be any hypergraph with a special edge consisting of $N_\lambda - k\lfloor N_\lambda/k \rfloor$ vertices. In particular, $H_{\mathrm{PC}}(n, p\,; k)$ is not $k$-uniform when $N_\lambda$ is not divisible by $k$. Therefore, as long as properties of $k$-uniform hypergraphs are concerned, we do not have to describe $H_{\mathrm{PC}}(n, p\,; k)$ when $N_\lambda$ is not divisible by $k$.

We show that the Poisson cloning model $H_{\mathrm{PC}}(n, p\,; k)$, $k \geq 2$, is contiguous to the classical model $H(n, p\,; k)$ when the expected average degree is a constant.

**Theorem 1.1** (restated). *Suppose $k \geq 2$ and $p = \Theta(n^{1-k})$. Then for any collection $\mathcal{H}$ of $k$-uniform simple hypergraphs,*

$$c_1 \Pr[H_{\mathrm{PC}}(n, p\,; k) \in \mathcal{H}] \leq \Pr[H(n, p\,; k) \in \mathcal{H}] \leq c_2\big(\Pr[H_{\mathrm{PC}}(n, p\,; k) \in \mathcal{H}]^{\frac{1}{k}} + e^{-n}\big),$$

*where*

$$c_1 = k^{1/2} e^{\frac{p}{n}\binom{k}{2}\binom{n}{k} + \frac{p^2}{2}\binom{n}{k}} + O(n^{-1/2}), \quad c_2 = \Big(\frac{k}{k-1}\Big)\Big(c_1(k-1)\Big)^{1/k} + o(1),$$

*and $o(1)$ goes to 0 as $n$ goes to infinity.*

*Proof.* See [30]. □

## 3. The λ-cell and the cut-off line algorithm

To generate a uniform random perfect matching of $N_\lambda$ clones, we may keep matching $k$ unmatched clones uniformly at random (cf. Example 2.1). Another way is to choose the first clone as we like and match it to $k-1$ clones selected uniformly at random among all other unmatched clones (cf. Example 2.2). As there are many ways to choose the first clone, we may take a way that makes the given problem easier to analyze. Formally, a sequence $\mathcal{S} = (S_i)$ of choice functions determines how to choose the first clone at each step, where $S_i$ tells us which unmatched clone is to be the first clone for the $i^{\text{th}}$ edge in the random perfect matching. A choice function may be deterministic or random. If less than $k$ clones remain unmatched, the edge consisting of those clones will be added. The clone chosen by $S_i$ is called the $i^{\text{th}}$ chosen clone, or simply a chosen clone.

We also present a more specific way to select the $k-1$ random clones to be matched to the chosen clone. The way introduced here will be useful to solve problems mentioned in the introduction. First, independently assign to each clone a uniform

random real number between 0 and $\lambda = p\binom{n-1}{k-1}$. For the sake of convenience, a clone is called the largest, the smallest, etc. if so is the number assigned to it. In addition, a clone is called $\theta\lambda$-*large* (resp. $\theta\lambda$-small) if its assigned number is larger than or equal to (resp. smaller than) $\theta\lambda$. To visualize the labelled clones with assigned numbers, one may consider $n$ horizontal line segments from $(0, j)$ to $(\lambda, j)$, $j = 0, \ldots, n-1$ in the two-dimensional plane $\mathbb{R}^2$. The $v_j$-clone with assigned number $x$ can be regarded as the point $(x, j)$ in the corresponding line segment. Then each line segment with the points corresponding to clones with assigned numbers is an independent Poisson arrival process with density 1, up to time $\lambda$. The set of these Poisson arrival processes is called a *Poisson* $(\lambda, n)$*-cell* or simply a $\lambda$*-cell*.

We will consider sequences of choice functions that choose an unmatched clone without changing the joint distribution of the numbers assigned to all other unmatched clones. Such a choice function is called oblivious. A sequence of oblivious choice functions is also called oblivious. The choice function that chooses the largest unmatched clone is not oblivious, as the numbers assigned to the other clones must be smaller than the largest assigned number. As an example of an oblivious choice function one may consider the choice function that chooses a $v$-clone for a vertex $v$ with fewer than 3 unmatched clones. For a more general example, let a vertex $v$ and its clones be called $t$*-light* if there are fewer than $t$ unmatched $v$-clones.

**Example 3.1.** Suppose there is an order of all clones which is independent of the assigned numbers. The sequence of the choice functions that choose the first $t$-light clone is oblivious.

A cut-off line algorithm is determined by an oblivious sequence of choice functions. Once a clone is obliviously chosen, the largest $k-1$ clones among all unmatched clones are to be matched to the chosen clone. This may be further implemented by moving the cut-off line to the left until $k-1$ vertices are found: Initially, the cut-off line of the $\lambda$-cell is the vertical line in $\mathbb{R}^2$ containing the point $(\lambda, 0)$. The initial cut-off value, or cut-off number, is $\lambda$. At the first step, once the chosen clone is given, move the cut-off line to the left until exactly $k-1$ unmatched clones, excluding the chosen clone, are on or in the right side of the line. The new cut-off value, which is denoted by $\Lambda_1$, is to be the number assigned to the $(k-1)^{\text{th}}$ largest clone. The new cut-off line is, of course, the vertical line containing $(\Lambda_1, 0)$. Repeating this procedure, one may obtain the $i^{\text{th}}$ cut-off value $\Lambda_i$ and the corresponding cut-off line.

Notice that, after the $i^{\text{th}}$ step ends with the cut-off value $\Lambda_i$, all numbers assigned to unmatched clones are i.i.d. uniform random numbers between 0 to $\Lambda_i$, as the choice functions are oblivious. Let $N_i$ be the number of unmatched clones after step $i$. That is, $N_i = N_\lambda - ik$. Since the $(i+1)^{\text{th}}$ choice function tells how to choose the first clone to form the $(i+1)^{\text{th}}$ edge without changing the distribution of the assigned numbers, the distribution of $\Lambda_{i+1}$ is the distribution of the $(k-1)^{\text{th}}$ largest number among $N_i - 1$ independent uniform random numbers between 0 and $\Lambda_i$. Let $1 - T_j$ be the random variable representing the largest number among $j$ independent uniform random numbers between 0 and 1. Or equivalently, $T_j$ is the random variable

representing the smallest number among the random numbers. Then the largest number among the $N_i - 1$ random numbers has the same distribution as $\Lambda_i(1 - T_{N_i-1})$. Repeating this $k-1$ times, we have

$$\Lambda_{i+1} = \Lambda_i(1 - T_{N_i-1})(1 - T_{N_i-2}) \dots (1 - T_{N_i-k+1}),$$

and hence

$$\begin{aligned}\Lambda_{i+1} &= \Lambda_i(1 - T_{N_i-1}) \dots (1 - T_{N_i-k+1}) \\ &= \Lambda_{i-1}(1 - T_{N_{i-1}-1}) \dots (1 - T_{N_{i-1}-k+1}) \cdot (1 - T_{N_i-1}) \dots (1 - T_{N_i-k+1}) \\ &= \lambda \prod_{\substack{j=N_\lambda-1 \\ k\nmid N_\lambda-j}}^{N_\lambda-(i+1)k+1} (1 - T_j).\end{aligned}$$

It is crucial to observe that, once $N_\lambda$ is given, all $T_i$ are mutually independent random variables. This makes the random variable $\Lambda_i$ highly concentrated near its mean, which enables us to develop theories as if $\Lambda_i$ were a constant. The cut-off value $\Lambda_i$ will provide enough information to resolve some otherwise difficult problems.

For $\theta$ in the range $0 \le \theta \le 1$, let $\Lambda(\theta)$ be the cut-off value when $\big(1 - \theta^{\frac{k}{k-1}}\big)\lambda n$ or more clones are matched for the first time. Conversely, let $N(\theta)$ be the number of matched clones until the cut-off line reaches $\theta\lambda$.

**Lemma 3.2** (Cut-off line lemma). *Let $k \ge 2$ and $\lambda > 0$ be fixed. Then for $\theta_1 < 1$ uniformly bounded below from 0 and $0 < \Delta \le n$,*

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le1} |\Lambda(\theta) - \theta\lambda| \ge \tfrac{\Delta}{n}\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}$$

*and*

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le1} |N(\theta) - (1 - \theta^{\frac{k}{k-1}})\lambda n| \ge \Delta\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}.$$

*Proof.* See [30]. □

## 4. Large deviation inequalities

In this section a generalized Chernoff bound and an inequality for random process is given. Let $X_1, \dots, X_m$ be a sequence of random variables such that the distribution of $X_i$ is determined if all the values of $X_1, \dots, X_{i-1}$ are known. For example, $X_i = \Lambda(\theta_i)$ with $1 \ge \theta_1 \ge \cdots \ge \theta_m \ge 0$ in a Poisson $\lambda$-cell. If the upper and/or lower bounds are known for the conditional means $E[X_i|X_1, \dots, X_{i-1}]$ and for the conditional second and third monuments, then Chernoff type large deviation inequalities may be obtained not only for $\sum_{j=1}^m X_j$ but for $\min_{1\le i\le m} \sum_{j=1}^i X_j$ and/or

$\max_{1\le i\le m}\sum_{j=1}^{i} X_j$. Large deviation inequalities for such minima or maxima are especially useful in various situations. Lemma 3.2 can be shown using such inequalities too.

**Lemma 4.1.** *Let $X_1, \dots, X_m$ be a sequence of random variables. Suppose that*

$$E[X_i|X_1, \dots, X_{i-1}] \le \mu_i, \tag{4.1}$$

*and that there are positive constants $a_i$ and $b_i$ such that*

$$E[(X_i - \mu_i)^2|X_1, \dots, X_{i-1}] \le a_i, \tag{4.2}$$

*and*

$$E[(X_i - \mu_i)^3 e^{\xi(X_i-\mu_i)}|X_1, \dots, X_{i-1}] \le b_i \quad \text{for all } 0 \le \xi \le \xi_0. \tag{4.3}$$

*Then for any $\alpha$ with $0 < \alpha \le \xi_0\big(\sum_{i=1}^m a_i\big)^{1/2}$,*

$$\Pr\Big[\sum_{i=1}^m X_i \ge \sum_{i=1}^m \mu_i + \alpha\Big(\sum_{i=1}^m a_i\Big)^{1/2}\Big] \le \exp\bigg(-\frac{\alpha^2}{2}\bigg(1 + \frac{\alpha\sum_{i=1}^m b_i}{3\big(\sum_{i=1}^m a_i\big)^{3/2}}\bigg)\bigg).$$

*Similarly,*

$$E[X_i|X_1, \dots, X_{i-1}] \ge \mu_i \tag{4.4}$$

*together with* (4.2) *and*

$$E[(X_i - \mu_i)^3 e^{\xi(X_i-\mu_i)}|X_1, \dots, X_{i-1}] \ge b_i \quad \text{for all } \xi_0 \le \xi < 0 \tag{4.5}$$

*implies that*

$$\Pr\Big[\sum_{i=1}^m X_i \le \sum_{i=1}^m \mu_i - \alpha\Big(\sum_{i=1}^m a_i\Big)^{1/2}\Big] \le \exp\bigg(-\frac{\alpha^2}{2}\bigg(1 - \frac{\alpha\sum_{i=1}^m b_i}{3\big(\sum_{i=1}^m a_i\big)^{3/2}}\bigg)\bigg).$$

*Proof.* See [30]. □

As it is sometimes tedious to point out the value of $\alpha$ and to check the required bounds for it, the following forms of inequalities are often more convenient.

**Corollary 4.2** (Generalized Chernoff bound). *If $\delta\xi_0\sum b_i \le \sum a_i$ for some $0 < \delta \le 1$, then* (4.1)–(4.3) *imply*

$$\Pr\Big[\sum_{i=1}^m X_i \ge \sum_{i=1}^m \mu_i + R\Big] \le e^{-\frac{1}{3}\min\{\delta\xi_0 R,\ R^2/\sum_{i=1}^m a_i\}}$$

*for all $R > 0$. Similarly, if $-\delta\xi_0\sum b_i \le \sum a_i$ for some $0 < \delta \le 1$, then* (4.2), (4.4) *and* (4.5) *yield*

$$\Pr\Big[\sum_{i=1}^m X_i \le \sum_{i=1}^m \mu_i - R\Big] \le e^{-\frac{1}{3}\min\{\delta\xi_0 R,\ R^2/\sum_{i=1}^m a_i\}}$$

*for all $R > 0$.*

Let $X_\theta, \theta \geq 0$, be random variables which are possibly set-valued. Here $\theta$ may be integers as well as real numbers. Suppose that $\Gamma(\theta)$ is a random variable depending on $\{X_{\theta'}\}_{\theta' \leq \theta}$ and $\theta$, and

$$\psi = \psi(\{X_{\theta'}\}_{\theta' \leq \theta_1}; \theta_0, \theta_1) \quad \text{and} \quad \psi_\theta = \psi_\theta(\{X_{\theta'}\}_{\theta' \leq \theta_1}; \theta_0, \theta, \theta_1).$$

The random variables $\psi$ and $\psi_\theta$ are used to bound $\Gamma(\theta)$.

**Example 4.3.** Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli random variables with mean $p$ and $S_i = \sum_{j=1}^{i} X_j$. Set $\Gamma(i) = |S_i - ip|$ and

$$\psi = \Gamma(n) \quad \text{and} \quad \psi_i = |S_n - S_i - (n-i)p|.$$

Then, since

$$S_i - ip = S_n - np - (S_n - S_i - (n-i)p)$$

we have

$$\Gamma(i) \leq \psi + \psi_i.$$

**Example 4.4.** Consider the $(\lambda, n)$-cell defined in the previous section. Let $v_\theta$ be the vertex that has its largest clone at $(1-\theta)\lambda$. If such a vertex does not exist, $v_\theta$ is defined to be $\aleph$, assuming $\aleph \notin V$. As there is no possibility that two distinct clones are assigned the same number, $v_\theta$ is well-defined. Let $X_\theta = v_\theta$ and $V(\theta)$ be the set of vertices that contain no clone larger than or equal to $(1-\theta)\lambda$. That is, $V(\theta) = V \setminus \{v_{\theta'} : 0 \leq \theta' \leq \theta\}$. Clearly, $E[|V(\theta)|] = e^{-\theta\lambda} n$. Observing that for $\theta_0 \leq \theta \leq \theta_1$ one has

$$e^{-(\theta_1-\theta)\lambda}\big||V(\theta)| - e^{-\theta\lambda}n\big| \leq \big||V(\theta_1)| - e^{-\theta_1\lambda}n\big| + \big||V(\theta_1)| - e^{-(\theta_1-\theta)\lambda}|V(\theta)|\big|,$$

we may set $\Gamma(\theta) = ||V(\theta)| - e^{-\theta\lambda}n|$,

$$\psi = e^{(\theta_1-\theta_0)}\Gamma(\theta_1) \quad \text{and} \quad \psi_\theta = e^{(\theta_1-\theta_0)}\big||V(\theta_1)| - e^{-(\theta_1-\theta)\lambda}|V(\theta)|\big|.$$

We bound the probabilities $\max_{\theta_0 \leq \theta \leq \theta_1} \Gamma(\theta) \geq R$ and $\min_{\theta_0 \leq \theta \leq \theta_1} \Gamma(\theta) \leq R$ under some conditions.

**Lemma 4.5.** *Let $0 \leq \theta_0 < \theta_1$, $R = R_1 + R_2$, $R_1, R_2 > 0$ and $\Phi_\theta$ be events depending on $\{X_{\theta'}\}_{\theta' \leq \theta}$. If*

$$\Gamma(\theta) \leq \psi + \psi_\theta \quad \textit{for all } \theta_0 \leq \theta \leq \theta_1,$$

*then*

$$\Pr\Big[\max_{\theta_0 \leq \theta \leq \theta_1} \Gamma(\theta) \geq R\Big] \leq \Pr\big[\psi \geq R_1\big] + \Pr\Big[\bigcup_{\theta:\theta_0 \leq \theta \leq \theta_1} \overline{\Phi}_\theta\Big] + \max_{\theta:\theta_0 \leq \theta \leq \theta_1} \max_{\{X_{\theta'}\}_{\theta' \leq \theta}} 1(\Phi_\theta) \Pr\big[\psi_\theta \geq R_2 \mid \{X_{\theta'}\}_{\theta' \leq \theta}\big].$$

*Similarly, if*

$$\Gamma(\theta) \geq \psi + \psi_\theta \quad \textit{for all } \theta_0 \leq \theta \leq \theta_1,$$

*then*

$$\Pr\Big[\min_{\theta_0 \leq \theta \leq \theta_1} \Gamma(\theta) \leq -R\Big] \leq \Pr\big[\psi \leq -R_1\big] + \Pr\Big[\bigcup_{\theta:\theta_0 \leq \theta \leq \theta_1} \overline{\Phi}_\theta\Big] + \max_{\theta:\theta_0 \leq \theta \leq \theta_1} \max_{\{X_{\theta'}\}_{\theta' \leq \theta}} 1(\Phi_\theta) \Pr\big[\psi_\theta \leq -R_2 \mid \{X_{\theta'}\}_{\theta' \leq \theta}\big].$$

*Proof.* See [30]. □

**Example 4.3** (continued). As

$$\Pr[\psi \geq R_1] \leq e^{-\Omega(\min\{R_1, \frac{R_1^2}{p(1-p)n}\})}$$

and

$$\Pr[\psi_i \geq R_2 | X_1, \ldots, X_i] = \Pr[\psi_i \geq R_2] \leq e^{-\Omega(\min\{R_2, \frac{R_2^2}{p(1-p)(n-i)}\})},$$

Lemma 4.5 for $R_1 = R_2 = R/2$ and $\Phi_\theta = \emptyset$ gives

$$\Pr[\max_{i:0 \leq i \leq n} |S_i - pi| \geq R] \leq e^{-\Omega(\min\{R, \frac{R^2}{p(1-p)n}\})}.$$

**Example 4.4** (continued). Since

$$|V(\theta)| = \sum_{v \in V} 1(v \text{ has no } (1-\theta)\lambda\text{-large clone})$$

is a sum of i.i.d. Bernoulli random variables with mean $e^{-\theta\lambda}$,

$$\Pr\big[\big||V(\theta)| - e^{-\theta\lambda}n\big| \geq R\big] \leq e^{-\Omega(\min\{R, \frac{R^2}{\theta n}\})},$$

especially

$$\Pr\big[\psi \geq R/2\big] \leq e^{-\Omega(\min\{R, \frac{R^2}{\theta_1 n}\})}.$$

Once $\{X_{\theta'}\}_{\theta' \leq \theta}$ is given, $V(\theta)$ is determined and

$$V(\theta_1) = \sum_{v \in V(\theta)} 1(v \text{ has no } (1-\theta_1)\lambda\text{-large clone})$$

is a sum of i.i.d. Bernoulli random variables with mean $e^{-(\theta_1-\theta)\lambda}$. Thus

$$\Pr\big[\psi_\theta \geq R/2 \mid \{X_{\theta'}\}_{\theta' \leq \theta}\big] \leq 2e^{-\Omega(\min\{R, \frac{R^2}{(\theta_1-\theta)|V(\theta)|}\})} \leq 2e^{-\Omega(\min\{R, \frac{R^2}{\theta n}\})},$$

and Lemma 4.5 for $\theta_0 = 0$ and $\Phi_\theta = \emptyset$ yields

$$\Pr\Big[\max_{\theta:0 \leq \theta \leq \theta_1} |V(\theta) - e^{-\theta\lambda}n| \geq R\Big] \leq 2e^{-\Omega(\min\{R, \frac{R^2}{\theta n}\})}.$$

## 5. Generalized cores and the main lemma

In this section we introduce generalized cores and the main lemma. The main lemma will play a crucial roles in the proofs of the theorems mentioned in the introduction.

We start with some terminology. A *generalized degree* is an ordered pair $(d_1, d_2)$ of non-negative integers. The inequality between two generalized degrees is determined by the inequality between the first coordinates and the reverse inequality between the second coordinates. That is, $(d_1, d_2) \geq (d_1', d_2')$ if and only if $d_1 \geq d_1'$ and $d_2 \leq d_2'$. A *property* for generalized degrees is simply a set of generalized degrees. A property $P$ is *increasing* if generalized degrees larger than an element in $P$ are also in $P$. When a property $P$ depends only on the first coordinate of generalized degrees, it is simply a property of degrees. For the $t$-core problem, we will use $P_{t\text{-core}} = \{(d_1, d_2) : d_1 \geq t\}$. To estimate the size of the largest component, we will set $P_{\text{comp}} = \{(d_1, d_2) : d_2 = 0\}$.

Given the Poisson $\lambda$-cell on the set $V$ of $n$ vertices and $\theta$ with $0 \leq \theta \leq 1$, let $d_v(\theta)$ be the number of $v$-clones smaller than $\theta\lambda$. Similarly, $\bar{d}_v(\theta)$ is the number of $v$-clones larger than or equal to $\theta\lambda$. Then $D_v(\theta) := (d_v(\theta), \bar{d}_v(\theta))$ are i.i.d. random variables. In particular, for any property $P$ the events $D_v(\theta) \in P$ are independent and occur with the same probability, say $p(\theta, \lambda; P)$, or simply $p(\theta)$.

For an increasing property $P$, the $P$-process is defined as follows. Construct the Poisson $\lambda$-cell as described in Section 3, where $\lambda = p\binom{n-1}{k-1}$. The vertex set $V = \{v_0, \ldots, v_{n-1}\}$ will be regarded as an ordered set so that the $i^{\text{th}}$ vertex is $v_{i-1}$. The $P$-process is a generalization of Example 2.2 for which choice functions choose $t$-light clones.

**The $P$-process.** Initially, the cut-off value $\Lambda = \lambda$. Activate all vertices $v$ with $D_v(1) \notin P$. All clones of the activated vertices are activated too. Put those clones in a stack in an arbitrary order. However, this does not mean that the clones are removed from the $\lambda$-cell.

(a) If the stack is empty, go to (b). If the stack is nonempty, choose the first clone in the stack and move the cut-off line to the left until the largest $k-1$ unmatched clones, excluding the chosen clone, are found. (So, the cut-off value $\Lambda$ keeps decreasing.) Then match the $k-1$ clones to the chosen clone. Remove all matched clones from the stack and repeat the process. A vertex that has not been activated is to be activated as soon as $D_v(\Lambda/\lambda) \notin P$. This can be done even before all $k-1$ clones are found. Its unmatched clones are to be activated too and put into the stack immediately. Clones found while moving the cut-off line are also in the stack until they are matched.

(b) Activate the first vertex in $V$ which has not been activated. Its clones are activated too. Put those clones into the stack. Then go to (a).

Clones in the stack are called active. The steps carried out by the instruction described in (b) are called forced steps as it is necessary to artificially activate a vertex.

When the cut-off line is at $\theta\lambda$, all $\theta\lambda$-large clones are matched or will be matched at the end of the step and all vertices $v$ with $D_v(\theta) \notin P$ have been activated. All other

vertices can have been activated only by forced steps. Let $V(\theta) = V_P(\theta)$ be the set of vertices $v$ with $D_v(\theta) \in P$, and let $M(\theta) = M_P(\theta)$ be the number of $\theta\lambda$-large clones plus the number of $\theta\lambda$-small clones of vertices $v$ not in $V(\theta)$. That is,

$$M(\theta) = \sum_{v\in V} \bar{d}_v(\theta) + d_v(\theta)1(v \notin V(\theta)) = \sum_{v\in V} \bar{d}_v(\theta) + d_v(\theta)1(D_v(\theta) \notin P).$$

Recalling that $N(\theta)$ is the number of matched clones until the cut-off line reaches $\theta\lambda$, the number $A(\theta)$ of active clones (when the cut-off value $\Lambda$ is) at $\theta\lambda$ is at least as large as $M(\theta) - N(\theta)$. On the other hand, the difference $A(\theta) - (M(\theta) - N(\theta))$ is at most the number $F(\theta)$ of clones activated in forced steps until $\theta\lambda$, i.e.,

$$M(\theta) - N(\theta) \le A(\theta) \le M(\theta) - N(\theta) + F(\theta). \tag{5.1}$$

As the cut-off lemma gives a concentration inequality for $N(\theta)$,

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le 1} |N(\theta) - (1 - \theta^{\frac{k}{k-1}})\lambda n| \ge \Delta\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\}},$$

a concentration inequality for $M(\theta)$ will be enough to obtain a similar inequality for $B(\theta) := M(\theta) - N(\theta)$. More precisely, we will show that under appropriate hypotheses

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le 1} |M(\theta) - (\lambda - q(\theta))n| \le \Delta\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})},$$

where

$$q(\theta) = q(\theta, \lambda; P) = E\big[d_v(\theta)1(D_v(\theta) \in P)\big].$$

As the $d_v(\theta)$'s and $D_v(\theta)$'s are identically distributed, $q(\theta)$ does not depend on $v$. Also, recall that $p(\theta) = \Pr[D_v(\theta) \in P]$.

As we will see later, $B(\theta)$ is very close to $A(\theta)$. Hence a concentration inequality for $B(\theta)$ plays a very important roles in all of our proofs.

**Lemma 5.1** (Main lemma). *In the $P$-process, if $\theta_1 < 1$ uniformly bounded from below by 0, $1 - p(\theta_1) = O(1 - \theta_1)$ and $p(\theta_1) = \Omega(1)$, then for all $\Delta$ in the range $0 < \Delta \le n$ we have*

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le 1} \big||V(\theta)| - p(\theta)n\big| \le \Delta\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}$$

*and*

$$\Pr\Big[\max_{\theta:\theta_1\le\theta\le 1} \big|B(\theta) - (\lambda\theta^{\frac{k}{k-1}} - q(\theta))n\big| \le \Delta\Big] \le 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}.$$

*Proof.* See [30]. □

## 6. Cores of random hypergraphs

In this section we prove Theorem 1.7. Let $\lambda > 0$ and $H(\lambda) = H_{\mathrm{PC}}(n, p)$, where $\lambda = p\binom{n-1}{r-1}$. Let the property $P = \{(d_1, d_2) : d_1 \geq t\}$. Then

$$p(\theta) = Q(\theta\lambda, t) \quad \text{and} \quad q(\theta) = \theta\lambda Q(\theta\lambda, t-1).$$

The main lemma gives

**Corollary 6.1.** *For $\theta_1 \leq 1$ uniformly bounded from below by 0 and $\Delta$ in the range $0 < \Delta \leq n$,*

$$\Pr\Big[\max_{\theta:\theta_1\leq\theta\leq 1} \big||V(\theta)| - Q(\theta\lambda, t)n\big| \geq \Delta\Big] \leq 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{n}\})}$$

*and*

$$\Pr\Big[\max_{\theta:\theta_1\leq\theta\leq 1} \big|B(\theta) - (\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1))\theta\lambda n\big| \geq \Delta\Big] \leq 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{n}\})}.$$

*Subcritical Region*: For $\lambda = \lambda_{\mathrm{crt}} - \sigma$, $\sigma > 0$ and $\theta_1 = \delta/\lambda_{\mathrm{crt}}$ with $\delta = 0.1$ it is easy to see that there is a constant $c > 0$ such that

$$(\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1))\theta\lambda n \geq c\sigma n \quad \text{for all } \theta \text{ with } \theta_1 \leq \theta \leq 1.$$

Let $\tau$ be the first time the number $A(\theta)$ of active clones at $\theta\lambda$ becomes 0. Then the second part of Corollary 6.1 gives

$$\begin{aligned}\Pr[\tau \geq \theta_1] &\leq \Pr[B(\theta) = 0 \text{ for some } \theta \text{ with } \theta_1 \leq \theta \leq 1] \\ &\leq \Pr\Big[\max_{\theta:\theta_1\leq\theta\leq 1} \big|B(\theta) - (\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1))\theta\lambda n\big| \geq c\sigma n\Big] \\ &\leq 2e^{-\Omega(\sigma^2 n)}.\end{aligned}$$

As $\theta_1\lambda \leq \theta_1\lambda_{\mathrm{crt}} = \delta$, and hence $Q(\theta_1\lambda, t) \leq \delta/2$ for $t \geq 2$, the first part of Corollary 6.1 yields

$$\Pr[|V_t(H_{\mathrm{PC}}(n, p\,; k))| \geq \delta n] \leq \Pr[\tau \geq \theta_1] + \Pr[|V(\theta_1)| \geq \delta n] \leq 2e^{-\Omega(\sigma^2 n)}.$$

Therefore Theorem 1.1 implies that

$$\Pr[|V_t(H(n, p\,; k))| \geq \delta n] \leq 2e^{-\Omega(\sigma^2 n)}.$$

To complete the proof, we observe that the $t$-core of size $i$ has at least $ti/k$ edges. Let $Z_i$ be the number of subgraphs on $i$ vertices with at least $ti/k$ edges, $i = i_0, \ldots, \delta n$, where $i_0 = i_0(k, t)$ is the least $i$ such that $\binom{i}{k} \geq ti/k$. Then

$$E[Z_i] \leq \binom{n}{i}\binom{\binom{i}{k}}{ti/k} p^{ti/k} \leq \frac{n^i}{i!}\frac{i^{ti}}{(ti/k)!}p^{ti/k} =: L_i, \tag{6.1}$$

where $ti/k$ actually means $\lceil ti/k \rceil$. Hence

$$\frac{L_{i+k}}{L_i} = O\left(\frac{n^k}{i^k}\frac{i^{kt}}{i^t}n^{-(k-1)t}\right) = O\left(\left(\frac{i}{n}\right)^{(k-1)t-k}\right) = O(\delta^{(k-1)(t-1)-1}).$$

That is, $L_{i+k}/L_i$ exponentially decreases. For $i = i_0, \ldots, i_0 + k - 1$,

$$L_i = O(n^i n^{-i(k-1)t/k}) = O(n^{-i(t-1-t/k)})$$

implies that

$$\Pr[V_t(H(n, p; k)) \neq \emptyset] \leq 2e^{-\Omega(\sigma^2 n)} + O(n^{-i_0(t-1-t/k)}),$$

as desired. □

*Supercritical region*: We will prove the following theorem.

**Theorem 6.2.** *Suppose that* $p\binom{n-1}{k-1} \geq \lambda_{\mathrm{crt}} + \sigma$ *and* $0 < \delta \leq 1$. *Then, with probability* $1 - 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}$, $V_t = V_t(H_{\mathrm{PC}}(n, p; k))$ *satisfies*

$$Q(\theta_\lambda \lambda, t)n - \delta n \leq |V_t| \leq Q(\theta_\lambda \lambda, t)n + \delta n, \tag{6.2}$$

*and the degrees of vertices of the* $t$*-core are i.i.d.* $t$*-truncated Poisson random variables with parameter* $\Lambda_t := \theta_\lambda \lambda + \beta$ *for some* $\beta$ *with* $|\beta| \leq \delta$. *Moreover, the distribution of the* $t$*-core is the same as that of the* $t$*-truncated Poisson cloning model with parameters* $|V_t|$ *and* $\Lambda_t$.

*Proof.* Let $\lambda = \lambda_{\mathrm{crt}} + \sigma$, $\sigma > 0$ and $\theta_\lambda$ be the largest solution for the equation

$$\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1) = 0.$$

Then it is not hard to check that there are constants $c_1, c_2 > 0$ such that for $\theta$ in the range $\theta_\lambda \leq \theta \leq 1$,

$$\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1) \geq c_1\sigma^{1/2}(\theta - \theta_\lambda),$$

and for $\theta$ in the range $\theta_\lambda - c_2\sigma^{1/2} \leq \theta \leq \theta_\lambda$,

$$\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1) \leq -c_1\sigma^{1/2}(\theta_\lambda - \theta).$$

Let $\tau$ be the largest $\theta$ with $A(\theta) = 0$. Then $V(\tau)$ is the $t$-core of $H_{\mathrm{PC}}(n, p; k)$. For $\theta_1 = \theta_\lambda + \delta$ and $\theta_2 = \theta_\lambda - \min\{\delta, c_2\sigma^{1/2}\}$ with $0 < \delta \leq 1$, Corollary 6.1 gives

$$\begin{aligned}\Pr[\tau \geq \theta_1] &\leq \Pr[B(\theta) = 0 \text{ for some } \theta \text{ with } \theta_1 \leq \theta \leq 1] \\ &\leq \Pr\Big[\max_{\theta:\theta_1\leq\theta\leq 1} |B(\theta) - (\theta^{\frac{1}{k-1}} - Q(\theta\lambda, t-1))\theta\lambda n| \geq c_1\sigma^{1/2}\delta n\Big] \\ &\leq 2e^{-\Omega(\delta^2\sigma n)}\end{aligned}$$

and

$$\begin{aligned}\Pr[\tau < \theta_2] &\le \Pr[B(\theta_2) > 0] \\ &\le \Pr\left[|B(\theta_2) - (\theta_2^{\frac{1}{k-1}} - Q(\theta_2\lambda, t-1))\theta_2\lambda n| \ge c_1\sigma^{1/2}\min\{\delta, c_2\sigma^{1/2}\}n\right] \\ &\le 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}.\end{aligned}$$

Since $\frac{d}{d\theta}Q(\theta\lambda, t) = \lambda P(\theta\lambda, t-1) \le \lambda$, we have

$$Q(\theta_1\lambda, t) \le Q(\theta_\lambda\lambda, t) + \lambda\delta, \quad \text{and} \quad Q(\theta_2\lambda, t) \ge Q(\theta_\lambda\lambda, t) - \lambda\delta,$$

and Corollary 6.1 implies that

$$\Pr[V(\theta_1) - Q(\theta_\lambda\lambda, t)n \ge 2\lambda\delta n] \le 2e^{-\Omega(\delta^2 n)}$$

and

$$\Pr[V(\theta_2) - Q(\theta_\lambda\lambda, t)n \le -2\lambda\delta n] \le 2e^{-\Omega(\delta^2 n)}.$$

Therefore

$$\Pr[|\tau - \theta_\lambda| > \delta] \le \Pr[\tau \ge \theta_1] + \Pr[\tau \le \theta_2] \le 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}$$

and, replacing $\delta$ by $\frac{\delta}{2\lambda}$,

$$\begin{aligned}\Pr[|V(\tau) - Q(\theta_\lambda\lambda, t)n| \ge \delta n] &\le \Pr[\tau \ge \theta_1] + \Pr[\tau \le \theta_2] + 2e^{-\Omega(\delta^2 n)} \\ &\le 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}.\end{aligned}$$

Clearly, once $V(\tau)$ and $\Lambda_t := \tau\lambda$ are given, the residual degrees $d_v(\tau)$, $v \in V(\tau)$, are i.i.d. $t$-truncated Poisson random variables with parameter $\Lambda_t$. □

Once $V_t$ and $\Lambda_t$ are given, $|V_t(i)|$, $i \ge t$, is the sum of i.i.d. Bernoulli random variables with mean $p_i(\Lambda_t) := \frac{P(\Lambda_t, i)}{Q(\Lambda_t, t)}$. Similarly, the size of $W_t(i) = \bigcup_{j \ge i} V_t(j)$ is the sum of i.i.d. Bernoulli random variables with mean $q_i(\Lambda_t) := \frac{Q(\Lambda_t, i)}{Q(\Lambda_t, t)}$. Applying the generalized Chernoff bound (Lemma 4.2), we have

$$\Pr\left[\left||V_t(i)| - p_i(\Lambda_t)|V_t|\right| \ge \delta|V_t||V_t, \Lambda_t\right] \le 2e^{-\Omega(\delta^2|V_t|)}$$

and

$$\Pr\left[\left||W_t(i)| - q_i(\Lambda_t)|V_t|\right| \ge \delta|V_t||V_t, \Lambda_t\right] \le 2e^{-\Omega(\delta^2|V_t|)}.$$

Combining this with Lemma 6.2 and using

$$|P(\rho, i) - P(\rho', i)| \le |\rho - \rho'|, \quad \text{and} \quad |Q(\rho, i) - Q(\rho', i)| \le |\rho - \rho'|,$$

we obtain, for any $i$,

$$\Pr\left[\left||V_t(i)| - P(\theta_\lambda\lambda, i)n\right| \ge \delta n\right] \le 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})},$$

and

$$\Pr\big[\big||W_t(i)| - Q(\theta_\lambda\lambda, i)n\big| \geq \delta n\big] \leq 2e^{-\Omega(\min\{\delta^2\sigma n, \sigma^2 n\})}.$$

In particular, as $\theta_\lambda = \theta_{\rm crt} + \Theta(\sigma^{1/2})$ for uniformly bounded $\sigma$ it follows that for $\lambda = \lambda_{\rm crt} + \sigma$,

$$|V_t(i)| = (1 + O(\sigma^{1/2}))^i P(\theta_{\rm crt}\lambda_{\rm crt}, i)n + O\big((n/\sigma)^{1/2}\log n\big),$$

with probability $1 - 2e^{-\Omega(\min\{\log^2 n, \sigma^2 n\})}$.

## 7. The emergence of the giant component

In this section we just give ideas for the proof of Theorem 1.4. Let the property $P$ be $\{(d_1, d_2) : d_2 = 0\}$. Then $p(\theta) = e^{-(1-\theta)\lambda}$ and $q(\theta) = \theta\lambda e^{-(1-\theta)\lambda}$, and the main lemma gives

**Corollary 7.1.** *For $\theta_1 \geq 1$ uniformly bounded from above by 1 and $\Delta$ in the range $0 < \Delta \leq n$,*

$$\Pr\Big[\max_{\theta:0\leq\theta\leq\theta_1}\big||V(\theta)| - e^{-(1-\theta)\lambda}n\big| \geq \Delta\Big] \leq 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}$$

*and*

$$\Pr\Big[\max_{\theta:0\leq\theta\leq\theta_1}\big|B(\theta) - (\theta - e^{-(1-\theta)\lambda})\theta\lambda n\big| \geq \Delta\Big] \leq 2e^{-\Omega(\min\{\Delta, \frac{\Delta^2}{(1-\theta_1)n}\})}.$$

To estimate $A(\theta)$ it is now enough for us to estimate $F(\theta)$ by (5.1). Once good estimations for $F(\theta)$ are established, we may take similar (but slightly more complicated) approaches used in the previous section.

It is convenient to consider an (imaginary) secondary stack with parameter $\rho$, or simply $\rho$-secondary stack. Initially, the secondary stack with parameter $\rho$ consists of the first $\rho n$ vertices $v_0, \ldots, v_{\rho n-1}$ of $V$. The set of those $\rho n$ vertices is denoted by $V_\rho$. Whenever the primary stack is empty, the first vertex in the secondary stack that has not been activated must be activated. Its clones are activated too and put into the primary stack. The activated vertex as well as vertices activated by other means are no longer in the secondary stack. If the secondary stack is empty, go back to the regular procedure. This does not change the $P$-process at all, but will be used just for the analysis. Let $\tau_\rho$ be the largest $\tau$ such that, at $\tau\lambda$, the primary stack becomes empty after the secondary stack is empty. Thus, once the cut-off line reaches $\tau_\rho\lambda$, no active clones are provided from the secondary stack. Denote by $C(\rho)$ the union of the components containing any vertex in $V_\rho$.

The following lemma is useful to predict how large $\tau_\rho$ is.

**Lemma 7.2.** *Suppose* $0 < \delta,\ \rho < 1$ *and* $\theta_1, \theta_2 \le 1$ *are uniformly bounded from below by* 0. *Then*

$$\Pr[\tau_\rho \ge \theta_1] \le \Pr\Big[\min_{\theta:\theta_1\le\theta\le 1} B(\theta) \le -(1-\delta)\theta_1\lambda e^{-(1-\theta_1)\lambda}\rho n\Big] + 2e^{-\Omega(\delta^2\rho n)},$$

*and conversely,*

$$\Pr[\tau_\rho \le \theta_2] \le \Pr[B(\theta_2) \ge -(1+\delta)\theta_2\lambda e^{-(1-\theta_2)\lambda}\rho n] + 2e^{-\Omega(\delta^2\rho n)}.$$

*Proof.* See [30]. □

Once the value of $\tau_\rho$ is known quite precisely, a good estimation of $F(\theta)$ is possible. Using similar (but slightly more complicated) arguments used in the previous section, estimation of $A(\theta)$ is also possible. Due to space limitation, the proof of Theorem 1.4 is omitted.

## 8. Closing remarks

The Poisson $\lambda$-cell is introduced to analyze those properties of $G_{\mathrm{PC}}(n, p)$, for which the degrees are i.i.d. Poisson random variables with mean $\lambda = p(n-1)$. Then various nice properties of Poisson random variables are used to analyze sizes of the largest component and the $t$-core of $G_{\mathrm{PC}}(n, p)$. We believe that the approaches presented in this paper are useful to analyze problems with similar flavors, especially problems related to branching processes. For example, we can easily modify the proofs of Theorem 1.7 to analyze the pure literal rule for the random $k$-SAT problems, $k \ge 3$. Another example may be the Karp–Sipser Algorithm to find a large matching of the random graph. (See [29], [3].) In a subsequent paper, we will analyze the structure of the 2-core of $G(n, p)$ and the largest strong component of the random directed graph as well as the pure literal rule for the random 2-SAT problem.

For the random (hyper)graph with a given sequence $(d_i)$, we may also introduce the $(d_i)$-cell, in which the vertex $v_i$ has $d_i$ clones and each clone is assigned a uniform random real number between 0 and the average degree $\frac{1}{n}\sum_{i=0}^{n-1} d_i$. Though it is not possible to use all of the nice properties of Poisson random variables any more, we believe that the $(d_i)$-cell equipped the cut-off line algorithm can be used to prove stronger results for the $t$-core problems considered in various papers including [13], [21], [22], [26], [37].

Recall that the degrees in $G(n, p)$ has the binomial distribution with parameters $n-1$ and $p$. By introducing the Poisson cloning model, we somehow first take the limit of the binomial distribution, which is the Poisson distribution. In general, many limiting distributions like Poisson and Gaussian ones have nice properties. In our opinion this is because various small differences are eliminated by taking the limits, and limiting distributions have some symmetric and/or invariant properties. Thus

one may wonder whether there is an infinite graph that shares most properties of the random graphs $G(n, p)$ with large enough $n$. So, in a sense, the infinite graph, if it exists, can be regarded as the limit of $G(n, p)$. An infinite graph which Aldous [1] considered to solve the linear assignment problem may or may not be a (primitive) version of such an infinity graph. Though it may be impossible to construct such a graph, the approaches taken in this paper might be useful to find one, if any.

**Acknowledgement.** The author thanks C. Borgs, J. Chayes, B. Bollobás and Y. Peres for helpful discussions.

## References

[1] Aldous, D., The zeta(2) limit in the random assignment problem. *Random Structures Algorithms* **18** (2001), 381–418.

[2] Alon, N., Spencer, J., *The Probabilistic Method.* 2nd ed., Wiley-Interscience, New York, NY, 2000.

[3] Aronson, J., Frieze, A., Pittel, B., Maximum matchings in sparse random graphs: Karp-Sipser revisited. *Random Structures Algorithms* **12** (1998), 111–178.

[4] Athreya, K. B., Ney, P. E., *Branching processes*. Grundlehren Math. Wiss. 196, Springer-Verlag, Berlin 1972.

[5] Bollobás, B., The evolution of random graphs. *Trans. Amer. Mat. Soc.* **286** (1984), 257–274.

[6] Bollobás, B., The evolution of sparse graphs. In *Graph Theory and Combinatorics* (ed. by B. Bollobás), Academic Press, London 1984, 35–57.

[7] Bollobás, B., *Random graphs*. Academic Press, London 1985.

[8] Bollobás, B., The chromatic number of random graphs. *Combinatorica* **8** (1988), 49–56.

[9] Bollobás, B., Erdős, P., Cliques in random graphs. *Math. Proc. Cambridge Philos. Soc.* **80** (1976), 419–427.

[10] Bollobás, B., Frieze A. M., On matchings and Hamiltonian cycles. In *Random Graphs '83* (ed. by M. Karoński and A. Ruciński), North-Holland, Amsterdam, New York 1985, 23–46.

[11] Bollobás, B., Thomason A., Random graphs of small order. In *Random Graphs '83* (ed. by M. Karoński and A. Ruciński), North-Holland, Amsterdam, New York 1985, 47–97.

[12] Chvátal, V., Almost all graphs with 1.44 edges are 3-colorable. *Random Structures Algorithms* **2** (1991), 11–28.

[13] Cooper, C., The cores of random hypergraphs with a given degree sequence. *Random Structures Algorithms* **25** (2004), 353–375.

[14] Erdős, P., Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.* **53** (1947), 292–294.

[15] Erdős, P., Rényi, A., On random graphs I. *Publ. Math. Debrecen* **6** (1959), 290–297.

[16] Erdős, P., Rényi, A., On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960), 17–61.

[17] Erdős, P., Rényi, A., On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hung.* **12** (1961), 261–267.

[18] Erdős, P., Rényi, A., On random matrices. *Publ. Math. Inst. Hung. Acad. Sci.* **8** (1964), 455–461.

[19] Erdős, P., Rényi, A., On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hung.* **17** (1966), 359–368.

[20] Erdős, P., Rényi, A., On random matrices II. *Stud. Sci. Math. Hung.* **3** (1968), 459–464.

[21] Fernholz D., Ramachandran V., The giant $k$-core of a random graph with a specified degree sequence. Manuscript, 2003.

[22] Fernholz D., Ramachandran V., Cores and connectivity in sparse random graphs. Tech. Report TR-04-13, University of Texas at Austin, Department of Computer Science, Austin, 2004.

[23] Flajolet, D., Knuth D. E., Pittel, B., The first cycle in an evolving graph. *Discrete Math.* **75** (1989), 167–215.

[24] Janson, S., Poisson convergence and Poisson processes with applications to random graphs. *Stochastic Process. Appl.* **26** (1988), 1–30.

[25] Janson, S., Knuth D. E., Łuczak, T., Pittel, B., The birth of the giant component. *Random Structures Algorithms* **3** (1993), 233–358.

[26] Janson, S., Łuczak, M. J., A simple solution to the $k$-core problem. Manuscript, 2005.

[27] Janson, S., Łuczak, T., Ruciński, A., *Random graphs*. Wiley-Interscience, New York, NY, 2000.

[28] Karp, R. M., The transitive closure of a random digraph. *Random Structures Algorithms* **1** (1990), 73–93.

[29] Karp, R. M., Sipser, M., Maximum matchings in sparse random graphs. In *Proceedings of the 22nd IEEE Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1981, 364–375.

[30] Kim., J.-H., Poisson cloning model for random graph. Manuscript, 2005.

[31] Komlś, J., Szemerédi E., Limit distributions for the existence of Hamilton cycles in a random graph. *Discrete Math.* **43** (1983), 55–63.

[32] Łuczak, T., Component behavior near the critical point of the random graph process. *Random Structures Algorithms* **1** (1990), 287–310.

[33] Łuczak, T., The chromatic number of random graphs. *Combinatorica* **11** (1991), 45–54.

[34] Łuczak, T., Size and connectivity of the $k$-core of a random graph. *Discrete Math.* **91** (1991), 61–68.

[35] Łuczak, T., Pittel, B., Wierman, J., The birth of the giant component. *Trans. Amer. Mat. Soc.* **341** (1994), 721–748.

[36] Matula, D., The largest clique size in a random graph. *Tech. Rep., Dept. Comp. Sci.*, Southern Methodist University, Dallas, 1976.

[37] Molloy, M., Cores in random random hypergraphs and Boolean formulas. *Random Structures Algorithms* **27** (2005), 124–135.

[38] Pittel, B., Spencer J., Wormald, N., Sudden emergence of a giant $k$-core in a random graph. *J. Combin. Theory Ser. B* **67** (1996), 111–151.

[39] Shamir, E., Spencer, J., Sharp concentration of the chromatic number on random graph $G_{n,p}$. *Combinatorica* **7** (1987), 124–129.

Microsoft Research, One Microsoft Way, Redmond, WA 98052, U.S.A

E-mail: jehkim@microsoft.com

# Randomness and regularity

Tomasz Łuczak

**Abstract.** For the last ten years the theory of random structures has been one of the most rapidly evolving fields of discrete mathematics. The existence of sparse structures with good 'global' properties has been vital for many applications in mathematics and computer science, and studying properties of such objects led to many challenging mathematical problems. In the paper we report on recent progress on this subject related to some variants of Szemerédi's Regularity Lemma.

**Mathematics Subject Classification (2000).** Primary 05C80, 05D05; Secondary 05C35, 05C65, 05D40.

**Keywords.** Random graphs, extremal graph theory, regularity lemma, removal lemma, density theorems.

## 1. Introduction

In the paper we consider 'extremal' properties of families of sets, i.e., we study the size of maximal subfamilies of a given family with certain property. Let $[A]^r$ be the family of all $r$-sets (i.e. sets of $r$ elements) contained in $A$; if $A = [n] = \{1, 2, \ldots, n\}$ we put $[n]^r = [[n]]^r$. Two classical examples of extremal results for $[n]^r$ are Szemerédi's and Turán's theorems. Let us recall that Szemerédi's density theorem [17] states that if $r_k(n)$ denote the maximum size of a subset of $[n] = [n]^1$ which contains no non-trivial arithmetic progression of length $k$, then, $r_k(n) = o(n)$. In order to formulate Turán's theorem, we need some notation. For given $r, s, n$, $n \geq s > r \geq 2$, let $\mathrm{ex}([s]^r, [n]^r)$ denote the size of the largest family $\mathcal{A} \subseteq [n]^r$ such that for no set $B \subseteq [n]$, $|B| = s$, we have $[B]^r \subseteq \mathcal{A}$. Furthermore, let

$$\alpha(m, r) = \limsup_{n\to\infty} \frac{\mathrm{ex}([m]^r, [n]^r)}{\binom{n}{r}} \,. \tag{1}$$

Turán's theorem [21] states that $\alpha(m, 2) = \frac{m-2}{m-1}$ for $m \geq 2$. Let us remark that we do not know the value of $\alpha(m, r)$ for any pair $(m, r)$ with $m > r > 2$; e.g., the question whether $\alpha(4, 3) = 5/9$ is a well known open problem of extremal set theory.

The main problem we are concerned in this paper is the existence of families of $r$-sets which are 'sparse', or, at least, 'locally sparse', yet preserve some of the properties of $[n]^r$ stated in the theorems above. In the following section, we state a few specific problems on the existence of locally sparse structures with good 'global'

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

properties. Then, we explain why the standard probabilistic method cannot be directly used to study extremal properties of graphs and hypergraphs. Next we recall another important result of modern combinatorics: Szemerédi's Regularity Lemma and show how it could help in dealing with such problems. We conclude with a few remarks on possible generalizations of known results and some speculation on developments which are still to come.

## 2. Locally sparse structures with good extremal properties

Let us introduce first some notation. An $r$-uniform hypergraph is a pair $H = (V, E)$, where $V$ is the set of vertices of $H$ and $E \subseteq [V]^r$ denotes the set of its edges. We say that a hypergraph $H' = (V', E')$ is a subhypergraph of a hypergraph $H'' = (V'', E'')$, if $V' \subseteq V''$ and $E' \subseteq E''$. The complete $r$-uniform hypergraph $([m], [m]^r)$ we denote by $K_m^{(r)}$, and set $K_m^{(2)} = K_m$. A 2-uniform hypergraph is called a graph.

Let $H = (V, E)$ be an $r$-uniform hypergraph, and let $C = \{W_1, \ldots, W_t\}$ be a family of $s$-subsets of $V$ such that $[W_i]^r \subseteq E$, for $i = 1, \ldots, , t$. We say that $C$ is a loose $(s, t)$-circuit if $t \geq 3$, $(s, t) \neq (3, 3)$, $W_i \cap W_{i+1} \neq \emptyset$, for $i = 1, 2, \ldots, t-1$, and $W_1 \cap W_t \neq \emptyset$. We call $C$ a tight $(s, t)$-circuit if either $t = 2$ and $|W_1 \cap W_2| \geq r+1$, or $t \geq 3$ and $|W_i \cap W_{i+1}| \geq r$, for $i = 1, 2, \ldots, t-1$, as well as $|W_1 \cap W_t| \geq r$.

Finally, for $r$-uniform hypergraphs $H = (V, E)$ and $H' = (V', E')$, let $\mathrm{ex}(H', H)$ be the number of edges in the largest subhypergraph of $H$ which contains no copies of $H'$, and $\overline{\mathrm{ex}}(H', H) = \mathrm{ex}(H', H)/|E|$. It is easy to see (e.g., [8], Prop.8.4.), that for a given $H'$ the function $\overline{\mathrm{ex}}(H', H)$ is maximized for complete $H$, i.e., $\overline{\mathrm{ex}}(H', H) \leq \overline{\mathrm{ex}}(H', [V]^r)$.

One of the first results on the existence of locally sparse structures with good extremal properties was proved by Erdős [1] nearly fifty years ago. It states that there are graphs with large girth and no large independent sets (and so with large chromatic number).

**Theorem 2.1.** *For each $\ell$ and $\varepsilon > 0$ there exists a graph $G(\ell, \varepsilon) = (V, E)$ such that $G(\ell, \varepsilon)$ contains no $(2, t)$-circuits with $t \leq \ell$, but each subset $W \subseteq V$ such that $|W| \geq \varepsilon |V|$ contains an edge of $G(\ell, \varepsilon)$.*

In the following section we present Erdős' elegant non-constructive proof of this fact. Then we shall try to use a similar idea to get the following sparse version of Turan's theorem.

**Conjecture 2.2.** For any $r$, $s$, $\varepsilon > 0$, and $\ell$, there exist an $n = n(r, m, \ell, \varepsilon)$ and an $r$-uniform hypergraph $G^{(r)}(s, \ell, \varepsilon) = (V, E)$ such that

(i) $G^{(r)}(s, \ell, \varepsilon)$ contains no tight $(s, t)$-cycles with $2 \leq t \leq \ell$;

(ii) each subhypergraph $H^{(r)} \subseteq G^{(r)}(s, \ell, \varepsilon)$ with at least $(\alpha(s, r) + \varepsilon)|E|$ edges contains a subset $B$, $|B| = s$, such that $[B]^r \subseteq H^{(r)}$, i.e.,

$$\overline{\mathrm{ex}}([n]^s, G^{(r)}(s, \ell, \varepsilon)) \leq \alpha(s, r) + \varepsilon\,.$$

In Sections 3-5 below we describe how to approach Conjecture 2.2 using a special version of the Regularity Lemma. Here we remark only that the existence of $G^{(r)}(s, \ell, \varepsilon)$ has been shown only for $r = 2, s = 3$ (Frankl and Rödl [2] and Haxell *et al.* [7]), $r = 2, s = 4$ (Kohayakawa *et al.* [10]), and recently for $r = 2, s = 5$ (Gerke *et al.* [4]).

We conclude this section with a conjecture on a sparse version of Szemerédi's density theorem. Here a $(k, t)$-arithmetic circuit is a family of $t$ non-trivial arithmetic progressions $A_1, \ldots, A_t$ of length $k$ such that $A_i \cap A_{i+1} \neq \emptyset$ for $i = 1, 2 \ldots, t-1$, and $A_1 \cap A_t \neq \emptyset$.

**Conjecture 2.3.** For any $k$, $\ell$, and $\alpha > 0$, there exist an $\varepsilon = \varepsilon(\alpha, k, \ell) > 0$, $n = n(k, \alpha, \ell)$, and a set $A = A(k, \ell, \alpha, n) \subseteq [n]$ such that

(i) $A$ contains no $(k, t)$-arithmetic circuits for $t \leq \ell$;

(ii) any non-trivial arithmetic progression of length $n^{\varepsilon}$ in $[n]$ contains at most $k$ elements of $A$;

(iii) each subset $B$ of $A$ with at least $\alpha|A|$ elements contains a non-trivial arithmetic progression of length $k$.

Kohayakawa *et al.* [9] showed the existence of $A = A(3, \ell, \alpha, n)$ for any $\alpha > 0$ and $\ell$. Their proof was based on the idea used by Ruzsa and Szemerédi [16] to show that $r_3(n) = o(n)$. Let us also mention that since Szemerédi's density theorem can be deduced from some extremal results for hypergraphs (see Frankl and Rödl [3], Nagle *et al.* [13], and Rödl and Skokan [15]) it is in principle possible, although somewhat unlikely, that one can imitate the argument from [9] and verify Conjecture 2.3 for all $k \geq 4$ (cf., Conjecture 6.2 below).

## 3. Random structures

For $0 \leq p \leq 1$ and natural numbers $n, r$, let $\mathbb{G}^{(r)}(n, p)$ denote the random $r$-uniform hypergraph with vertex set $[n]$, where edges $\mathbb{G}^{(r)}(n, p)$ are chosen from $[n]^r$ independently with probability $p$. Thus, the number of edges of $\mathbb{G}^{(r)}(n, p)$ is a binomially distributed random variable with parameters $\binom{n}{r}$ and $p$. Typically, we are interested only in the asymptotic behavior of $\mathbb{G}^{(r)}(n, p)$ when $n \to \infty$ and the probability $p$ may depend on $n$. In particular, we say that for a given function $p = p(n)$ the hypergraph $\mathbb{G}^{(r)}(n, p)$ has a property $\mathcal{A}$ a.a.s. if the probability that $\mathbb{G}^{(r)}(n, p)$ has $\mathcal{A}$ tends to 1 as $n \to \infty$. Since in this note we deal mainly with graphs, instead of $\mathbb{G}^{(2)}(n, p)$ we write briefly $\mathbb{G}(n, p)$.

Let us recall Erdős' proof of Theorem 2.1. Fix $\ell$ and $\varepsilon > 0$. Let $n$ be very large and $p$ be the probability which is neither too small (so $\mathbb{G}(n, p)$ contains no large independent sets) nor too large (so $\mathbb{G}(n, p)$ is locally sparse). A good choice for $p$ is, say, $p = p(n) = n^{-1+1/2\ell}$, but, for $n$ large enough, any $p = p(n)$ such that $10/\varepsilon \leq np \leq n^{1/\ell}/10$ would do.

Let $X = X(n, \ell)$ be the random variable which counts $(2, t)$-circuits with $t \leq \ell$ in $\mathbb{G}(n, p)$. Then, for $n$ large enough, we have

$$\mathbb{E}X \leq \sum_{i=3}^{\ell} \binom{i}{2}\binom{i}{2}^{2i} p^i \leq \ell^{2\ell+3}(np)^{\ell} \leq n^{1/2}. \tag{2}$$

Thus, from Markov's inequality, $\Pr(X \geq n/2) \leq 2n^{-1/2}$, and so, for large enough $n$, with probability at least $2/3 > 1 - 2n^{-1/2}$, we have $X \leq n/2$. On the other hand, for the number $Y = Y(n, k)$ of independent sets of size $k$ in $\mathbb{G}(n, p)$ we have

$$\Pr(Y > 0) \leq \mathbb{E}X = \binom{n}{k}(1-p)^{\binom{k}{2}} \leq 2^n \exp\left(-p\binom{k}{2}\right). \tag{3}$$

If $k = \varepsilon n/2$ then $\Pr(Y > 0)$ tends to 0 as $n \to \infty$, i.e., for $n$ large enough we have $\Pr(Y > 0) < 2/3$. Now, let $G(\ell, \varepsilon)$ be a graph obtained from $\mathbb{G}(n, p)$ by removing one vertex from each $(2, t)$-circuit with $t \leq \ell$. Then, with probability at least $1/3$, $G(\ell, \varepsilon)$ fulfills the assertion of Theorem 2.1.

The main goal of this paper is to discuss how one can verify Conjecture 2.2 using a modified version of Erdős' approach. We shall concentrate on the simplest non-trivial case of Conjecture 2.2, when $s = 3$ and $r = 2$. In order to deduce the existence of $G^{(2)}(3, \ell, \varepsilon)$ from appropriate properties of the random graph $\mathbb{G}(n, p)$ first we need to guess what value of $p$ we are to use. More specifically, we should find the smallest possible value of $p_0 = p_0(n)$ such that a.a.s. in each subgraph of $\mathbb{G}(n, p_0)$ which contains, say, 51% of its edges one can find a triangle. Note that if a graph $G$ contains $m$ edges and $t$ triangles, then there is a triangle-free subgraph of $G$ with at least $m - t$ edges. Thus, it seems that in $\mathbb{G}(n, p_0)$ the expected number of triangles (equal to $\binom{n}{3}p_0^3$) must be at least of the order of the expected number of edges (equal $\binom{n}{2}p_0$), i.e., $p_0 = p_0(n)$ should be at least as large as $\Omega(n^{-1/2})$. It turns out that this necessary condition is also sufficient and the following holds (see Frankl and Rödl [2], Haxell *et al.* [7]).

**Theorem 3.1.** *For every $\delta > 0$ there exists $c = c(\delta)$ such that if $p = p(n) \geq cn^{-1/2}$, then a.a.s. each subgraph of $\mathbb{G}(n, p)$ with at least $(1/2 + \delta)\binom{n}{2}p$ edges contains a triangle.*

Let us try to prove Theorem 3.1 using Erdős' argument. To this end one has to bound the expected number of triangle-free subgraphs $H$ of $\mathbb{G}(n, p)$, containing 51% of edges of $\mathbb{G}(n, p)$, using a formula similar to (3). In order to do that one needs to estimate the probability that such a large subgraph $H$ of $\mathbb{G}(n, p)$ contains no triangles. The first problem which immediately emerges is the fact that our argument must depend strongly on the fact that $H$ has more than 51% of edges of $\mathbb{G}(n, p)$, since in every graph $G$ one can find a large bipartite subgraph which contains more than half of its edges. Thus, we have to use some property of $H$ shared by all subgraphs of $\mathbb{G}(n, p)$ with more than half of its edges, and does not hold for, say, bipartite

subgraphs of $\mathbb{G}(n, p)$; i.e., we should consider only graphs $H$ which are 'essentially non-bipartite'. Then, we need to show that a.a.s. each subgraph of $\mathbb{G}(n, p)$ containing at least 51% of its edges is 'essentially non-bipartite', and estimate the probability that a 'random essentially non-bipartite' graph is triangle-free.

However, now we face another, more serious obstacle. The number of subsets of the set of edges of $\mathbb{G}(n, p)$ is much larger then the number of subsets of the set of vertices of $\mathbb{G}(n, p)$. Consequently, the factor $2^n$ in (3) should be replaced by $\exp(\Omega(n^2 p))$. Hence, we should estimate the probability that a 'random essentially non-bipartite' subgraph $H$ of $\mathbb{G}(n, p)$ is triangle-free by a quantity which is much smaller than the probability that $\mathbb{G}(n, p)$ contains no edges at all! This is the crucial and most difficult part of the whole argument. It is also precisely the reason why we can show Conjecture 2.2 only for $r = 2$ and $s = 3, 4, 5$; for all other cases the proof breaks at this point.

Finally, it is easy to check that if $p = p(n) = n^{-1/2+o(1)}$, then for any given $\ell$ a.a.s. $\mathbb{G}(n, p)$ contains fewer than $o(n^2 p)$ tight $(3, t)$-circuits for $t \leq \ell$ which can be removed from $\mathbb{G}(n, p)$ without affecting much its extremal properties. Unfortunately, one cannot deal in the same way with loose $(3, t)$-circuits. The reason is quite simple: for $t \geq 4$ the number of loose $(3, t)$-circuits grows much faster than the number of triangles, because, roughly speaking, two triangles of $\mathbb{G}(n, p)$ are much more likely to share a vertex than an edge. Clearly, the same is true if instead of $\mathbb{G}(n, p)$ we consider a shadow of $\mathbb{G}^{(3)}(n, p)$, i.e., we randomly generate triples of vertices and then replace each of them by a triangle. Still, it is not inconceivable that Conjecture 2.2 can be settled in the affirmative by a non-constructive method using more sophisticated models of random hypergraphs; there have been a fair amount of attempts in this direction but so far all of them have failed miserably.

## 4. Regularity Lemma

One of the main ingredients of Szemerédi's ingenious proof of the density theorem was the Regularity Lemma which for the last thirty years has become one of the most efficient tools of modern graph theory. In order to formulate it rigorously we need a few technical definitions.

For a graph $G = (V, E)$ and $W, W' \subseteq V$ let $e(W, W')$ denote the number of edges joining $W$ and $W'$. A pair $(A, B)$ of disjoint subsets of vertices of $G$ is called an $\varepsilon$-regular pair, if for every subsets $A' \subseteq A$, $|A'| \geq \varepsilon|A|$, $B' \subseteq B$, $|B'| \geq \varepsilon|B|$,

$$\left| \frac{e(A', B')}{|A'||B'|} - \frac{e(A, B)}{|A||B|} \right| \leq \varepsilon. \tag{4}$$

An $\varepsilon$-regular pair behaves in many respects as the bipartite random graph $\mathbb{G}(A, B, \rho)$, in which edges between $A$ and $B$ appear independently with probability $\rho = e(A, B)/|A||B|$. In particular, it is easy to check, that if a pair $(A, B)$

is $\varepsilon$-regular then the number of subgraphs of a given size in the bipartite subgraph $G[A, B]$ induced by $A \cup B$ in $G$ is close to the expected number of such subgraphs in $\mathbb{G}(A, B, \rho)$. For instance, if $(A, B)$ is $\varepsilon$-regular, then the number of cycles of length four in $G[A, B]$ is equal $(\rho^4/4 \pm h(\varepsilon))|A|^2|B|^2$, where $h(\varepsilon)$ is a function which tends to 0 as $\varepsilon \to 0$. The implication in the other direction holds as well: if the number of cycles of length four in $G[A, B]$ is smaller than $(\rho^4/4 + \varepsilon)|A|^2|B|^2$, then the pair $(A, B)$ is $h'(\varepsilon)$-regular for some function $h'(\varepsilon)$ which tend to 0 as $\varepsilon \to 0$. Let us also mention that $\varepsilon$-regularity implies the correct number of small subgraphs even if we consider more than one $\varepsilon$-regular pair. For instance, if three disjoint sets $A_1, A_2, A_3 \subseteq V$ are such that each of the pairs $(A_1, A_2)$, $(A_2, A_3)$, $(A_1, A_3)$ is $\varepsilon$-regular with density $\rho$, the number of triangles in the tripartite graph induced in $G$ by these sets is $(\rho^3 \pm h''(\varepsilon))|A_1||A_2||A_3|$, where $h''(\varepsilon) \to 0$ as $\varepsilon \to 0$.

A partition $V = V_1 \cup \cdots \cup V_k$ of the vertex set of a graph $G = (V, E)$ is called a $(k, \varepsilon)$-partition if for all $i, j = 1, 2, \ldots, k$ we have $\big||V_i| - |V_j|\big| \leq 1$ and all except at most $\varepsilon k^2$ pairs $(V_i, V_j)$, $1 \leq i < j \leq k$, are $\varepsilon$-regular. Now Szemerédi's Regularity Lemma (see [17] and [18]) can be stated as follows.

**Lemma 4.1** (Szemerédi's Regularity Lemma). *For every $\varepsilon > 0$ there exists $K$ such that each graph $G$ with more than $1/\varepsilon$ vertices admits a $(k, \varepsilon)$-regular partition for some $k$, $1/\varepsilon < k < K$.*

Note that if $k > 1/\varepsilon$, then for a $(k, \varepsilon)$-regular partition there are at most $n^2/k + \varepsilon n^2 \leq 2\varepsilon n^2$ edges of $G$ which are either contained inside sets $V_i$ or join pairs which are not $\varepsilon$-regular. Thus, Szemerédi's Regularity Lemma says that all but $2\varepsilon n^2$ edges of any graph $G$ can be partitioned into at most $k$ $\varepsilon$-regular pairs, for some $1/\varepsilon \leq k \leq K$, where $K$ does not depend on the number of vertices in $G$. Unfortunately, $K = K(\varepsilon)$ grows very fast to infinity as $\varepsilon \to 0$ (see Gowers [5]), so most of the applications of the Regularity Lemma give very poor bounds of estimated quantities.

The Regularity Lemma can be reformulated and generalized in several ways. For instance, one can view it as a statement on the compactness of certain metric space (Lovász and Szegedy [12]); an information-theoretic approach to it can be found in Tao [19]. Another versions of the Regularity Lemma ensure the existence of 'weak' $(k, \varepsilon)$-partitions, or 'partial $\varepsilon$-covers' consisting of reasonably large $\varepsilon$-pairs. However the two most important developments in this area are, in my opinion, generalizations of the Regularity Lemma to sparse graphs and to hypergraphs. In the following sections we discuss how the Regularity Lemma can be modified to work efficiently for sparse graphs; here we say a few words on a much harder (both to state and to prove) version of the Regularity Lemma for hypergraphs. Several years ago Frankl and Rödl [3] generalized the Regularity Lemma to $r$-uniform hypergraphs and proved it, together with a supplementary 'counting lemma', for $r = 3$. The case $r \geq 4$ has been dealt with by Rödl and Skokan [15] and Nagle *et al.* [13], and, independently, by Gowers [6]. As was noticed by Frankl and Rödl [3], their version of the Regularity Lemma implies the following Removal Lemma which, in turn, can be used to show Szemerédi's density theorem (for details see Rödl *et al.* [14]).

**Lemma 4.2** (Removal Lemma). *For $m \geq r \geq 2$ and every $\delta > 0$ there exist $\eta > 0$ and $n_0$ so that for every $r$-uniform hypergraph $F$ on $m$ vertices and $r$-uniform hypergraph $H$ on $n$, $n \geq n_0$, vertices the following holds. If $H$ contains at most $\eta n^m$ copies of $F$, then one can delete $\delta n^r$ edges of $H$ to destroy all copies of $F$.*

Thus, the Removal Lemma states that if the number of copies of $F$ is large enough, then they must be, in some sense, uniformly distributed in $H$, i.e., the large number of copies of $F$ makes a hypergraph $H$, or at least parts of it, close to a random graph. In fact all known proofs of Lemma 4.2 are based on this idea. We should apply the Regularity Lemma to $H$, and then show that, if the number of copies of $F$ in $H$ is large, then there exists a big random-like subgraph $H'$ of $H$ which contains an anticipated number of copies of $F$.

## 5. Regularity Lemma: sparse graphs

Note that Lemma 4.1 is basically meaningless for sparse graphs since the definition (4) of $\varepsilon$-regular pair $(A, B)$ does not say much on the distribution of edges between $A$ and $B$ if the density $\rho = e(A, B)/|A||B|$ is smaller than $\varepsilon$. Thus, let us modify the definition of an $\varepsilon$-regular pair by 'scaling' the density of the pair by $d$ which, typically, is the density of the graph $G = (V, E)$. Hence, we say that a pair $(A, B)$ of disjoint subsets of vertices of a graph $G = (V, E)$ is $(d, \varepsilon)$-regular, if for each pair of subsets $A' \subseteq A$, $|A'| \geq \varepsilon|A|$, $B' \subseteq B$, $|B'| \geq \varepsilon|B|$, we have

$$\left| \frac{e(A', B')}{|A'||B'|} - \frac{e(A, B)}{|A||B|} \right| \leq d\varepsilon. \tag{5}$$

If $d_G = |E|/\binom{|V|}{2}$ we call a $(d_G, \varepsilon)$-regular pair strongly $\varepsilon$-regular. A strongly $(k, \varepsilon)$-regular partition of vertices of $G$ is defined in a similar way as $(k, \varepsilon)$-regular partition. Moreover, we say that a graph $G = (V, E)$ with density $d_G = |E|/\binom{|V|}{2}$ is $(\eta, b)$-bounded if each subgraph $H$ of $G$ with $r \geq \eta|V|$ vertices contains not more than $br^2$ edges. As was observed independently by Kohayakawa and by Rödl (see Kohayakawa and Rödl [11] and references therein), one can mimic the proof of Szemerédi's Regularity Lemma to get the following result.

**Lemma 5.1.** *For every $\varepsilon > 0$ and $b$ there exist $\eta$ and $K$ such that each $(\eta, b)$-bounded graph $G$ with more than $1/\varepsilon$ vertices admits a strongly $(k, \varepsilon)$-regular partition for some $k$, $1/\varepsilon < k < K$.*

The assumption that $G$ is $(\eta, b)$-bounded is typically not very restrictive. For instance, if $\eta > 0$ and $np \to \infty$ as $n \to \infty$, then the random graph $\mathbb{G}(n, p)$ is a.a.s. $(\eta, 2)$-bounded. Consequently, a.a.s. each subgraph of such $\mathbb{G}(n, p)$ which contains at least half of its edges is $(\eta, 4)$-bounded.

A more serious problem is that, unlike in the dense case, from the fact that a sparse pair is strongly $\varepsilon$-regular it does not follow that the number of cycles of length four

in that pair is close to the number of cycles of length four in the random bipartite graph of the same density. In a similar way, for every $\varepsilon > 0$ there exists $\delta > 0$ and a tripartite graph $G$ with vertex set $V_1 \cup V_2 \cup V_3$, $|V_1| = |V_2| = |V_3| = n$ such that all three pairs $(V_1, V_2)$, $(V_2, V_3)$, and $(V_1, V_3)$ are strongly $\varepsilon$-regular pairs with densities larger than $\delta$, yet $G$ is triangle-free. Nevertheless, Kohayakawa, Łuczak, and Rödl conjectured in [10] that such triangle-free tripartite graphs consisting of dense $\varepsilon$-regular triples are so rare that a.a.s. the random graph $\mathbb{G}(n, p)$ contains none of them as a subgraph.

In order to state the conjecture rigorously we need one more definition. Let $\mathbb{G}(n, p; \varepsilon, s)$ be a graph chosen at random from the family of all $s$-partite graphs with vertex set $V_1 \cup V_2 \cup \cdots \cup V_s$, $|V_1| = |V_2| = \cdots = |V_s| = n$, such that for each $i$, $j$, $1 \le i < j \le s$, the pair $(V_i, V_j)$ spans a bipartite strongly $\varepsilon$-regular graph with $\lceil pn^2 \rceil$ edges. Then the conjecture of Kohayakawa, Łuczak and Rödl for complete graphs goes as follows (for a more general statement see [10]).

**Conjecture 5.2.** For every $s$ and $\delta > 0$ there exist $\varepsilon > 0$ and $C$ such that if $n^s p^{\binom{s}{2}} > Cn^2 p$, then the probability that $\mathbb{G}(n, p; \varepsilon, s)$ contains no copies of $K_s$ is smaller than $\delta^{n^2 p}$.

A stronger 'counting' version of Conjecture 5.2 goes as follows.

**Conjecture 5.3.** For every $s$ and $\delta > 0$ there exist $\varepsilon > 0$ and $C$ such that if $n^s p^{\binom{s}{2}} > Cn^2 p$, then the probability that $\mathbb{G}(n, p; \varepsilon, s)$ contains fewer than $n^s p^{\binom{s}{2}}/2$ copies of $K_s$ is smaller than $\delta^{n^2 p}$.

So far Conjectures 5.2 and 5.3 have been shown only for $s = 3, 4, 5$ (see Gerke *et al.* [4] and the references therein).

Let us observe that Theorem 3.1 follows immediately from the fact that Conjecture 5.2 holds for $s = 3$. Indeed, let us fix $\delta > 0$ and let $p = C/\sqrt{n}$, where $C$ is a large constant. Take a subgraph $H$ of $\mathbb{G}(n, p)$ with at least $(1/2 + \delta)\binom{n}{2} p$ edges. Choose $\varepsilon > 0$ much smaller than $\delta$ and apply Lemma 5.1 to $H$ to find in it a strong $(k, \varepsilon)$-partition with $1/\varepsilon < k < K$ (as we have already pointed out for every $\eta > 0$, a.a.s. $\mathbb{G}(n, p)$ is $(2, \eta)$-bounded and so $H$ is $(4, \eta)$-bounded and fulfills assumptions of the lemma). Since $H$ contains more than half of the edges of $\mathbb{G}(n, p)$, and edges in $\mathbb{G}(n, p)$ are uniformly distributed around the graph, there exist three sets $V'$, $V''$, $V'''$ of the partition such that each of the pairs $(V', V'')$, $(V'', V''')$, $(V', V''')$, is strongly $\varepsilon$-regular and has density at least $\delta p/10$. (Let us remark that now a vague notion of an 'essentially non-bipartite' subgraph $H$ we have used in Section 3 can be made precise: a graph $H$ is essentially non-bipartite if it contains a balanced tripartite graph on $\Omega(n)$ vertices which consists of three dense strongly $\varepsilon$-regular pairs.) Now, one can use Conjecture 5.2 and argue as in (3) that a.a.s. each tripartite subgraph of $\mathbb{G}(n, p)$ of such a type contains a triangle. Thus, $H$ contains a triangle and Theorem 3.1 follows.

Finally, let us also note that if, say, $p = \log n/\sqrt{n}$, then elementary calculations similar to that used by Erdős (cf. (3)) reveal that for every fixed $\ell$ a.a.s. the number

of tight $(3, t)$-circuits in $\mathbb{G}(n, p)$ with $t \leq \ell$ is $o(n^2 p)$. Thus, one can obtain a graph $G^{(2)}(3, \ell, \delta)$ with all the properties specified in Conjecture 2.2 by deleting from $\mathbb{G}(n, p)$ all edges which belong to tight $(3, t)$-circuits, $t \leq \ell$.

## 6. Final remarks

It is easy to see that, arguing as in the proof of Theorem 3.1 above, one can show the existence of a graph $G^{(2)}(s, \ell, \delta)$ (see Conjecture 2.2) for every $s$ for which Conjecture 5.2 holds. A precise formulation of analogs of Conjectures 5.2 and 5.3 for hypergraphs would become very technical, thus we only mention that if appropriately stated hypergraph version of Conjecture 5.3 is true then the following straightforward 'probabilistic' generalization of the Removal Lemma holds.

**Conjecture 6.1.** For $s > r \geq 2$ and every $\delta > 0$ there exist $\eta > 0$ such that a.a.s. in each subhypergraph $H$ of the random $r$-uniform hypergraph $\mathbb{G}^{(r)}(n, p)$ which contains fewer than $\eta n^s p^{\binom{s}{r}}$ copies of $K_s^{(r)}$ one can destroy all these copies by removing fewer than $\delta n^r p$ hyperedges of $H$.

An analogous question on the validity of a probabilistic version of Szemerédi's density theorem can be stated as follows.

**Conjecture 6.2.** For every $\delta > 0$ and $k$ there exists $\eta > 0$ such that a.a.s. in each subset $A$ of $\mathbb{G}^{(1)}(n, p)$ with fewer than $\eta n^2 p^k$ non-trivial arithmetic progressions of length $k$ all these progressions can be destroyed by removing fewer than $\delta np$ elements from $A$.

Conjecture 6.1 states that a.a.s. a random hypergraphs $\mathbb{G}^{(r)}(n, p)$ has a property $\mathcal{A}$ such that if a hypergraph $G$ has $\mathcal{A}$ each subgraph $H$ of $G$ with fewer than $\eta n^s p^{\binom{s}{r}}$ copies of $K_s^{(r)}$ can be made $K_s^{(r)}$-free by deleting fewer than $\delta n^r p$ hyperedges. One can ask if $\mathcal{A}$ follows from some simple property $\mathcal{A}'$, i.e., whether there is a compact characterization of 'pseudorandom' sparse hypergraphs. A natural candidate for $\mathcal{A}'$ is the property that the number of some special subhypergraphs in $G$ is close to the expected value of the number of such subhypergraphs in the random hypergraph with the same density. In the case of graphs a good choice for 'probing' graphs seem to be cycles of length four. It is known (see Thomasson [20]) that if the number of cycles of length four in a graph $G$ is close to the anticipated one, then edges in $G$ are 'uniformly distributed' around $G$. Nonetheless we do not know if the 'correct' number of cycles of length four, possibly matched with some additional requirements ensuring that $G$ is locally sparse, can guarantee that $G$ has good 'extremal' properties like those described in Conjecture 6.1. Another challenging problem is to strengthen the definition of a strongly $\varepsilon$-regular pair to, say, a 'super $\varepsilon$-regular pair' such that the analog of Lemma 5.1 remains valid in this setting (i.e., each dense subgraph of a random-like graph $G$ admits a 'super $(k, \varepsilon)$-partition') and furthermore, each

tripartite graph which consists of three dense super $\varepsilon$-regular pairs contains a triangle. Similar questions can be asked for hypergraphs, as well as for the subsets of $[n]$ (or, in somewhat more natural setting, for subsets of $\mathbb{Z}_n$, where $n$ is a prime).

## References


[1] Erdős, P. Graph theory and probability. *Canad. J. Math.* **11** (1959), 34–38.

[2] Frankl, P., and Rödl, V., Large triangle-free subgraphs in graphs without $K_4$. *Graphs Combin.* **2** (1986), 135–144.

[3] Frankl, P., and Rödl, V., Extremal problems on set systems. *Random Structures Algorithms* **20** (2002), 131–164.

[4] Gerke, S., Schickinger, T., Steger, A., $K_5$-free subgraphs of random graphs. *Random Structures Algorithms* **24** (2004), 194–232.

[5] Gowers, W. T., Lower bounds of tower type for Szemerédi's uniformity lemma. *Geom. Funct. Anal.* **7** (1997), 322–337.

[6] Gowers, W. T., Quasirandomness, counting and regularity for 3-uniform hypergraphs. *Combin. Probab. Comput.* **15** (2006), 143–184.

[7] Haxell, P. E., Kohayakawa, Y., Łuczak, T., Turán's extremal problem graphs: forbidding odd cycles. *Combinatorica* **16** (1996), 107–122.

[8] Janson, S., Łuczak, T., Ruciński, A., *Random Graphs*. Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York 2000.

[9] Kohayakawa, Y., Łuczak, T., and V. Rödl, Arithmetic progressions of length three in subsets of a random set. *Acta Arith.* **75** (1996), 133–163.

[10] Kohayakawa, Y., Łuczak, T., and Rödl, V., On $K_4$-free subgraphs of random graphs. *Combinatorica* **17** (1997), 173–213.

[11] Kohayakawa, Y., and Rödl, V., Regular pairs in sparse random graphs. *Random Structures Algorithms* **22** (2003), 359–434.

[12] Lovasz, L., Szegedy, B., Szemerédi's Lemma for the analyst. Preprint.

[13] Nagle, B., Rödl, V., Schacht, M., The counting lemma for regular k-uniform hypergraphs. *Random Structures Algorithms* **28** (2006), 113–179.

[14] Rödl, V., Nagle, B., Skokan, J., Schacht, M., Kohayakawa, Y., The hypergraph regularity method and its applications. *Proc. Nat. Acad. Sci. USA* **102** (2005) 8109–8113.

[15] Rödl, V., Skokan, J., Regularity lemma for k-uniform hypergraphs. *Random Structures Algorithms* **25** (2004), 1–42.

[16] Ruzsa, I., Szemerédi, E., Triple systems with no six points carrying three triangles. In *Combinatorics* (Proc. Fifth Hungarian Colloq. Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai 18, North-Holland, Amsterdam, New York 1978, 939–945.

[17] Szemerédi, E., On sets of integers containing no $k$ elements in arithmetic progression. *Acta Arith.* **28** (1975), 299–345.

[18] Szemerédi, E., Regular partitions of graphs. In *Problèmes Combinatoires et Théorie des Graphes* (ed. by J. Bermond et al.), Colloq. Internat. CNRS 260, CNRS, Paris 1978, 399–401.

[19] Tao, T., An information-theoretic proof of Szemerédi's regularity lemma. Preprint

[20] Thomason, A., Pseudorandom graphs. In *Random graphs '85* (Poznań, 1985), North-Holland Math. Stud. 144, North-Holland, Amsterdam 1987, 307–331.

[21] Turán, P., Egy gráfelméleti szélsöérték feladotról. *Mat. Fiz. Lapok* **48** (1941), 436–452; see also On the theory of graphs. *Colloq. Math.* **3** (1954), 19–30.

Adam Mickiewicz University, Faculty of Mathematics and CS, 61-614 Poznań, Poland
E-mail: tomasz@amu.edu.pl

# Additive combinatorics and geometry of numbers

Imre Z. Ruzsa*

**Abstract.** We meditate on the following questions. What are the best analogs of measure and dimension for discrete sets? How should a discrete analogue of the Brunn–Minkowski inequality look like? And back to the continuous case, are we happy with the usual concepts of measure and dimension for studying the addition of sets?

**Mathematics Subject Classification (2000).** Primary 11B75; Secondary 05-99, 11H06, 52C07.

**Keywords.** Sumsets, additive combinatorics, lattice points, volume.

## 1. Introduction

"Additive combinatorics" is a name coined by (I think) Tao and Van for the title of their book in preparation to denote the study of additive properties of general sets – mainly of integers, but also in other structures. Works on this topics are generally classified as additive or combinatorial number theory.

The first result that connects additive properties to geometrical position is perhaps the following theorem of Freiman.

**Theorem 1.1** (Freiman [3], Lemma 1.14). *Let $A \subset \mathbb{R}^d$ be a finite set, $|A| = m$. Assume that $A$ is proper $d$-dimensional, that is, it is not contained in any affine hyperplane. Then*

$$|A + A| \geq (d + 1)m - \frac{d(d + 1)}{2}.$$

This theorem is exact, equality can occur, namely it holds when $A$ is a "long simplex", a set of the form

$$L_{dm} = \{0, e_1, 2e_1, \ldots, (m - d)e_1, e_2, e_3, \ldots, e_d\}. \tag{1.1}$$

In particular, if no assumption is made on the dimension, then the minimal possible cardinality of the sumset is $2m - 1$, with equality for arithmetic progressions.

This result can be extended to sums of different sets. This extension is problematic from the beginning, namely the assumption "$d$-dimensional" can be interpreted in

*Supported by Hungarian National Foundation for Scientific Research (OTKA), Grants No. T 38396, T 43623, T 42750.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

different ways. We can stipulate that both sets be $d$-dimensional, or only one, or, in the weakest form, make this assumption on the sumset only.

An immediate extension of Freiman's above result goes as follows.

**Theorem 1.2** ([11], Corollary 1.1). *If $A, B \subset \mathbb{R}^d$, $|A| \le |B|$ and $\dim(A+B) = d$, then we have*

$$|A+B| \ge |B| + d|A| - \frac{d(d+1)}{2}.$$

We can compare these results to the continuous case. Let $A, B$ be Borel sets in $\mathbb{R}^d$; $\mu$ will denote the Lebesgue measure. The celebrated Brunn–Minkowski inequality asserts that

$$\mu(A+B)^{1/d} \ge \mu(A)^{1/d} + \mu(B)^{1/d}, \tag{1.2}$$

and here equality holds if $A$ and $B$ are homothetic convex sets, and under mild and natural assumptions this is the only case of equality. It can also be observed that the case $A = B$ is completely obvious here: we have

$$\mu(A+A) \ge \mu(2\cdot A) = 2^d \mu(A).$$

Also the constant $2^d$ is much larger than the constant $d+1$ in Theorem 1.1. This is necessary, as there are examples of equality, however, one feels that this is an exceptional phenomenon and better estimations should hold for "typical" sets. A further difference is the asymmetrical nature of the discrete result and the symmetry of the continuous one. Finally, when $|A|$ is fixed, Theorem 1.2 gives a linear increment, while (1.2) yields

$$\mu(A+B) \ge \mu(B) + d\mu(A)^{1/d}\mu(B)^{1-1/d}.$$

In the next section we tell what can be said if we use cardinality as the discrete analog of measure, and prescribe only the dimension of the sets. Later we try to find other spatial properties that may be used to study sumsets.

We meditate on the following questions (without being able to even conjecture a definitive answer). What are the best analogs of measure and dimension for discrete sets? How should a discrete analogue of the Brunn–Minkowski inequality look like? The partial answers also suggest questions in the continuous case. Should we be satisfied with the usual concepts of measure and dimension for studying the addition of sets?

Most of the paper is a survey, however, there are some new results in Sections 4 and 6.

We end the introduction by fixing some notations, which were tacitly used above.

For two sets $A, B$ (in any structure with an operation called addition) by their *sum* we mean the set

$$A+B = \{a+b : a \in A,\ b \in B\}.$$

We use $A - B$ similarly. For repeated addition we write

$$kA = A + \cdots + A \quad (k \text{ times}),$$

in contrast to

$$k \cdot A = \{ka : k \in A\}.$$

Mostly our sets will be in an Euclidean space $\mathbb{R}^d$, and $e_1, \dots, e_d$ will be the system of unit vectors. We define initially the *dimension* $\dim A$ of a set $A \subset \mathbb{R}^d$ as the dimension of the smallest affine hyperplane containing $A$. (This definition will be modified in Section 3).

## 2. Results using cardinality and dimension

We consider finite sets in an Euclidean space $\mathbb{R}^d$.

Put

$$F_d(m,n) = \min\{|A+B| : |A| = m, |B| = n,\ \dim(A+B) = d\},$$
$$F'_d(m,n) = \min\{|A+B| : |A| = m, |B| = n,\ \dim B = d\},$$
$$F''_d(m,n) = \min\{|A+B| : |A| = m, |B| = n,\ \dim A = \dim B = d\}.$$

$F_d$ is defined for $m+n \geq d+2$, $F'_d$ for $n \geq d+1$ and $F''_d$ for $m \geq d+1, n \geq d+1$. $F_d$ and $F''_d$ are obviously symmetric, while $F'_d$ may not be (and, in fact, we will see that for certain values of $m, n$ it is not), and they are connected by the obvious inequalities

$$F_d(m,n) \leq F'_d(m,n) \leq F''_d(m,n).$$

I determined the behaviour of $F_d$ and of $F'_d$ for $m \leq n$. The more difficult problem of describing $F''_d$ and $F'_d$ for $m > n$ was solved by Gardner and Gronchi [4]; we shall quote their results later.

To describe $F_d$ define another function $G_d$ as follows:

$$G_d(m,n) = n + \sum_{j=1}^{m-1} \min(d, n-j), \quad n \geq m \geq 1$$

and for $m > n$ extend it symmetrically, putting $G_d(m,n) = G_d(n,m)$. In other words, if $n - m \geq d$, then we have

$$G_d(m,n) = n + d(m-1).$$

If $0 \leq t = n - m < d$, then for $n > d$ we have

$$G_d(m,n) = n + d(m-1) - \frac{(d-t)(d-t-1)}{2} = n(d+1) - \frac{d(d+1)}{2} - \frac{t(t+1)}{2},$$

and for $n \leq d$

$$G_d(m,n) = n + \frac{(m-1)(2n-m)}{2}.$$

With this notation we have the following result.

**Theorem 2.1** ([11], Theorem 1). *For all positive integers $m$, $n$ and $d$ satisfying $m + n \geq d + 2$ we have*

$$F_d(m, n) \geq G_d(m, n).$$

Theorem 1.2 is an immediate consequence.

Theorem 2.1 is typically exact; the next theorem summarizes the cases when we have examples of equality.

**Theorem 2.2** ([11], Theorem 2). *Assume $1 \leq m \leq n$. We have*

$$F_d(m, n) = F'_d(m, n) = G_d(m, n)$$

*unless either $n < d + 1$ or $m \leq n - m \leq d$ (in this case $n \leq 2d$).*

The construction goes as follows.

Assume $1 \leq m \leq n$, $n \geq d + 1$. Let $B$ be a long simplex, $B = L_{dn}$ as defined in (1.1).

If $n - m \geq d$, we put

$$A = \{0e_1, 1e_1, \dots, (m-1)e_1\}.$$

This set satisfies $|A| = m$. The set $A + B$ consists of the vectors $ie_1$, $0 \leq i \leq n + m - d - 1$ and the vectors $ie_1 + e_j$, $0 \leq i \leq m - 1$, $2 \leq j \leq d$, consequently

$$|A + B| = n + d(m - 1) = G_d(m, n).$$

If $n - m = t < d$, write $t = d - k$ and assume $k \leq m$. Now $A$ is defined by

$$A = \{0e_1, 1e_1, \dots, (m-k)e_1\} \cup \{e_2, \dots, e_k\}.$$

This set satisfies $|A| = m$. The set $A+B$ consists of the vectors $ie_1, 0 \leq i \leq 2(n-d)$, the vectors $ie_1 + e_j$, $0 \leq i \leq n - d$, $2 \leq j \leq d$, finally $e_i + e_j$, $2 \leq i, j \leq k$, hence

$$\begin{aligned} |A + B| &= 2(n - d) + 1 + (d - 1)(n - d + 1) + \frac{k(k-1)}{2} \\ &= n(d + 1) - \frac{d(d+1)}{2} - \frac{t(t+1)}{2} = G_d(m, n). \end{aligned}$$

These constructions cover all pairs $m, n$ except those listed in Theorem 2.2. Observe that $A$ is also a long simplex of lower dimension. For a few small values the exact bounds are yet to be determined.

We now describe Gardner and Gronchi's [4] bound for $F'_d(m, n)$. Informally their main result (Theorem 5.1) asserts that the $|A + B|$ is minimalized when $B = L_{dn}$, a long simplex, and $A$ is as near to the set of points inside a homothetic simplex as possible. More exactly the define (for a fixed value of $n$) the weight of a point $x = (x_1, \dots, x_d)$ as

$$w(x) = \frac{x_1}{n - d} + x_2 + \cdots + x_d.$$

This defines an ordering by writing $x < y$ if either $w(x) < w(y)$ or $w(x) = w(y)$ and for some $j$ we have $x_j > y_j$ and $x_i = y_i$ for $i < j$.

Let $D_{dmn}$ be the collection of the first $m$ vectors with nonnegative integer coordinates in this ordering. We have $D_{dnn} = L_{dn} = B$, and, more generally, $D_{dmn} = rB$ for any integer $m$ such that

$$m = |rB| = (n-d)\binom{r+d-1}{d} + \binom{r+d-1}{d-1}.$$

For such values of $m$ we also have

$$|A+B| = |(r+1)B| = (n-d)\binom{r+d}{d} + \binom{r+d}{d-1}.$$

With this notation their result sounds as follows.

**Theorem 2.3** (Gardner and Gronchi [4], Theorem 5.1). *If $A, B \subset \mathbb{R}^d$, $|A| = m$, $|B| = n$ and $\dim B = d$, then we have*

$$|A+B| \geq |D_{dmn} + L_{dn}|.$$

For $m < n$ this reproves Theorem 2.2. For $m \geq n$ the extremal set $D_{dmn}$ is also $d$-dimensional, thus this result also gives the value of $F''_d$.

**Corollary 2.4.** *For $m \geq n > d$ we have*

$$F''_d(m,n) = F'_d(m,n) = |D_{dmn} + L_{dn}|.$$

A formula for the value of this function is given in [4], Section 6. We quote some interesting consequences.

**Theorem 2.5** (Gardner and Gronchi [4], Theorem 6.5). *If $A, B \subset \mathbb{R}^d$, $|A| = m \geq |B| = n$ and $\dim B = d$, then we have*

$$|A+B| \geq m + (d-1)n + (n-d)^{1-1/d}(m-d)^{1/d} - \frac{d(d-1)}{2}.$$

**Theorem 2.6** (Gardner and Gronchi [4], Theorem 6.6). *If $A, B \subset \mathbb{R}^d$, $|A| = m$, $|B| = n$ and $\dim B = d$, then we have*

$$|A+B|^{1/n} \geq m^{1/d} + \left(\frac{n-d}{d!}\right)^{1/d}.$$

This result is as close to the Brunn–Minkowski inequality as we can get by using only the cardinality of the summands.

## 3. The impact function and the hull volume

While we will focus our attention to sets in Euclidean spaces, some definitions and results can be formulated more clearly in a more general setting. So let now $G$ be a commutative group. For a fixed finite set $B \subset G$ we define its *impact function* by

$$\xi_B(m) = \xi_B(m, G) = \min\{|A + B| : A \subset G,\ |A| = m\}.$$

This is defined for all positive integers if $G$ is infinite, and for $m \le |G|$ if $G$ is finite.

This function embodies what can be told about cardinality of sumsets if one of the set is unrestricted up to cardinality. The name is a translation of Plünnecke's "Wirkungsfunktion", who first studied this concept systematically for density [9].

We will be interested mainly in the infinite case, and in this case the dependence on $G$ can be omitted.

**Lemma 3.1.** *Let $G$, $G'$ be infinite commutative groups, $G' \subset G$, and let $B \subset G'$ be a finite set. We have*

$$\xi_B(m, G) = \xi_B(m, G') \tag{3.1}$$

*for all $m$.*

*Proof.* Take an $A \subset G$, $|A| = m$ with $|A + B| = \xi_B(m, G)$. Let $A = A_1 \cup \dots \cup A_k$ be its decomposition according to cosets of $G'$. For each $1 \le i \le k$ take an element $x_i$ from the coset containing $A_i$ so that the sets $A_i - x_i$ are pairwise disjoint; this is easily done as long as $G'$ is infinite. The set

$$A' = \bigcup (A_i - x_i)$$

satisfies $A' \subset G'$, $|A'| = n$ and

$$|A' + B| \le \sum |A_i - x_i + B| = \sum |A_i + B| = |A + B| = \xi_B(m, G),$$

hence $\xi_B(m, G') \le \xi_B(m, G)$. The inequality in the other direction is obvious. □

In the case of finite groups the connection between $\xi_B(m, G)$ and $\xi_B(m, G')$ can also be described by arguments like in chapters 3 and 4 of Plünnecke's above mentioned book [9]. We restrict our attention to infinite groups, and henceforth omit the reference to $G$ and write just $\xi_B(m)$ instead.

Let $G$ be a torsionfree group. Take a finite $B \subset G$, and let $G'$ be the subgroup generated by $B - B$, that is, the smallest subgroup such that $B$ is contained in a single coset. Let $B' = B - a$ with some $a \in B$, so that $B' \subset G'$. The group $G'$, as any finitely generated torsionfree group, is isomorphic to the additive group $\mathbb{Z}^d$ for some $d$. Let $\varphi : G' \to \mathbb{Z}^d$ be such an isomorphism and $B'' = \varphi(B')$. By Lemma 3.1 we have

$$\xi_B = \xi_{B'} = \xi_{B''},$$

so when studying the impact function we can restrict our attention to sets in $\mathbb{Z}^d$ that contain the origin and generate the whole lattice; we then study the set "in its natural habitat".

**Definition 3.2.** Let $B$ be a finite set in a torsionfree group $G$. By the *dimension* of $B$ we mean the number $d$ defined above, and denote it by $\dim B$. By the *hull volume* of $B$ we mean the volume of the convex hull of the set $B''$ described above and denote it by $\mathrm{hv}\, B$.

The set $B''$ is determined up to an automorphism of $\mathbb{Z}^d$. These automorphisms are exactly linear maps of determinant $\pm 1$, hence the hull volume is uniquely defined.

Observe that this dimension is not the same as the dimension described in the Introduction; in the case when $B \subset \mathbb{R}^k$ with some $k$, this is its dimension over the field of rationals.

**Theorem 3.3.** *Let $B$ be a finite set in a torsionfree group $G$, $d = \dim B$, $v = \mathrm{hv}\, B$. We have*

$$\lim |kB| k^{-d} = v.$$

A proof can be found in [12], Section 11, though this form is not explicitly stated there. An outline is as follows. By using the arguments above we may assume that $B \subset \mathbb{Z}^d$, $0 \in B$ and $B$ generates $\mathbb{Z}^d$. Let $B^*$ be the convex hull of $B$. Then $kB$ is contained in $k \cdot B^*$. The number of lattice points in $k \cdot B$ is asymptotically $\mu(k \cdot B^*) = k^d v$; this yields an upper estimate. To get a lower estimate one proves that with some constant $p$, $kB$ contains all the lattice points inside translate of $(k-p) \cdot B^*$; this is Lemma 11.2 of [12].

This means that the hull volume can be defined without any reference to convexity and measure, and this definition can even be extended to commutative semigroups. This follows from the following result of Khovanskii [5], [6]; for a simple proof see [8].

**Theorem 3.4** (Khovanskii). *Let $B$ be a finite set in a commutative semigroup. There is a $k_0$, depending on the set $B$, such that $|kB|$ is a polynomial function of $k$ for $k > k_0$.*

**Definition 3.5.** Let $B$ be a finite set in a commutative semigroup, and let $vk^d$ be the leading term of the polynomial which coincides with $|kB|$ for large $k$. By the *dimension* of $B$ we mean the degree $d$ of this polynomial, and by the *hull volume* we mean the leading coefficient $v$.

It turns out that in $\mathbb{Z}^d$, hence in any torsionfree group, the dimension and hull volume determine the asymptotic behaviour of the impact function.

**Theorem 3.6.** *Let $B$ be a finite set in a torsionfree commutative group $G$, $d = \dim B$, $v = \mathrm{hv}\, B$. We have*

$$\lim(\xi_B(m)^{1/d} - m^{1/d}) = v^{1/d}.$$

This is the main result (Theorem 3.1) of [12]. In the same paper I announce the same result for non necessarily torsionfree commutative groups without proof (Theorem 3.4). In a general semigroup $A + B$ may consist of a single element, so an attempt to an immediate generalization fails.

**Problem 3.7.** Does the limit $\lim \xi_B(m)^{1/d} - m^{1/d}$ exist in general commutative semigroups? Is there a condition weaker than cancellativity to guarantee its positivity?

Theorem 3.6 can be effectivized as follows (Theorems 3.2 and 3.3 of [12]).

**Theorem 3.8.** *With the notations of the previous theorem, if $d \geq 2$ and $m \geq v$, we have*

$$\xi_B(m) \leq m + dv^{1/d}m^{1-1/d} + c_1 v^{2/d} m^{1-2/d},$$
$$\xi_B(m)^{1/d} - m^{1/d} \leq v^{1/d} + c_2 v^{2/d} m^{-1/d}.$$

*($c_1, c_2$ depend on $d$.) With $n = |B|$ for large $m$ we have*

$$\xi_B(m) \geq m + dv^{1/d}m^{1-1/d} - c_3 v^{\frac{d+3}{2d}} n^{-1/2} m^{1-\frac{3}{2d}},$$
$$\xi_B(m)^{1/d} - m^{1/d} \geq v^{1/d} - c_4 v^{\frac{d+3}{2d}} n^{-1/2} m^{-1/(2d)}.$$

Probably the real error terms are much smaller than these estimates. For $d = 1$ we have the obvious inequality $\xi_B(m) \leq m + v$, with equality for large $m$ because the integers $\xi_B(m) - m$ cannot converge to $v$ otherwise. For $d = 2$ already $\sqrt{\xi_B(m)} - \sqrt{m}$ can converge to $\sqrt{v}$ from both directions.

**Theorem 3.9.** *The impact function of the set $B = \{0, e_1, e_2\} \subset \mathbb{Z}^2$ satisfies*

$$\sqrt{\xi_B(m)} - \sqrt{m} > \sqrt{v} \tag{3.2}$$

*for all $m$.*

*The impact function of the set $B = \{0, e_1, e_2, -(e_1 + e_2)\} \subset \mathbb{Z}^2$ satisfies*

$$\sqrt{\xi_B(m)} - \sqrt{m} < \sqrt{v} \tag{3.3}$$

*for infinitely many $m$.*

Inequality (3.2) was announced in [12] without proof as Theorem 4.1, and it is a special case of Gardner and Gronchi's Theorem 2.6. Inequality 3.3 is Theorem 4.3 of [12].

I cannot decide whether there is a set such that $\sqrt{\xi_B(m)} - \sqrt{m} < \sqrt{v}$ for all $m$.

## 4. The impact volume

Besides cardinality we saw the hull volume as a contender for the title "discrete volume". For both we had something resembling the Brunn–Minkowski inequality;

for cardinality we had Gardner and Gronchi's Theorem 2.6, which has the (necessary) factor $d!$, and for the hull volume we have Theorem 3.6, which only holds asymptotically.

There is an easy way to find a quantity for which the analogue of the Brunn–Minkowski inequality hods exactly: we can make it a definition.

**Definition 4.1.** The $d$-dimensional *impact volume* of a set $B$ (in an arbitrarily commutative group) is the quantity

$$\mathrm{iv}_d(B) = \inf_{m\in\mathbb{N}} \left(\xi_B(m)^{1/d} - m^{1/d}\right)^d.$$

Note that the $d$ above may differ from the dimension of $B$, in fact, it need not be an integer. It seems, however, that the only really interesting case is $d = \dim B$.

The following statement list some immediate consequences of this definition.

**Statement 4.2.** *Let $B$ be a finite set in a commutative torsionfree group.*

(a) $\mathrm{iv}_d(B)$ *is a decreasing function of $d$.*

(b) *If $|B| = n$, then*

$$\mathrm{iv}_1(B) = n - 1$$

*and*

$$\mathrm{iv}_d(B) \le \left(n^{1/d} - 1\right)^d \tag{4.1}$$

*for every $d$.*

(c) $\mathrm{iv}_d(B) = 0$ *for $d > \dim B$.*

(d) *For every pair $A, B$ of finite sets in the same group and every $d$ we have*

$$\mathrm{iv}_d(A + B)^{1/d} \ge \mathrm{iv}_d(A)^{1/d} + \mathrm{iv}_d(B)^{1/d}. \tag{4.2}$$

The price we have to pay for the discrete Brunn–Minkowski inequality (4.2) is that there is no easy way to compute the impact volume for a general set. We have the following estimates.

**Theorem 4.3.** *Let $B$ be a finite set in a commutative torsionfree group,* $\dim B = d$, $|B| = n$. *We have*

$$\left(\frac{n-d}{d!}\right) \le \mathrm{iv}_d(B) \le \mathrm{hv}\, B, \tag{4.3}$$

*with equality in both places if $B$ is a long simplex.*

The first inequality follows form Theorem 2.6 of Gardner and Gronchi, the second from Theorem 3.6.

**Problem 4.4.** What is the *maximal* possible value of $\mathrm{iv}_d(B)$ for $n$-element $d$-dimensional sets? Is perhaps the bound in (4.1) exact?

We now describe the impact volume for another important class of sets, namely cubes.

**Theorem 4.5.** *Let $n_1, \dots, n_d$ be positive integers and let*

$$B = \{(x_1, \dots, x_d) \in \mathbb{Z}^d : 0 \le x_i \le n_i\}. \tag{4.4}$$

*We have*

$$\mathrm{iv}_d(B) = \mathrm{hv}\, B = v = n_1 \dots n_d.$$

**Problem 4.6.** Is is true that when $B$ is the set of lattice points within a convex lattice polytope, then hv $B$ and $\mathrm{iv}_d(B)$ are very near?

They may differ, as the second example in Theorem 3.9 shows.

We shall deduce Theorem 4.5 from the following one.

**Theorem 4.7.** *Let $G = G_1 \times G_2$ be a commutative group represented as the direct product of the groups $G_1$ and $G_2$. Let $B = B_1 \times B_2 \subset G$ be a finite set with $B_1 \subset G_1$, $B_2 \subset G_2$. We have*

$$\mathrm{iv}_d(B) \ge \mathrm{iv}_{d-1}(B_1)\mathrm{iv}_1(B_2). \tag{4.5}$$

*Proof.* Write $\mathrm{iv}_d(B) = v$, $\mathrm{iv}_{d-1}(B_1) = v_1$, $\mathrm{iv}_1(B_2) = v_2$ (which is $= |B| - 1$ if $G_2$ is torsionfree). We want to estimate $|A + B|$ from below for a general set $A \subset G$ with $|A| = m$.

Firs we transform them to some standard form; this will be the procedure what Gardner and Gronchi call compression. Let $A_1$ be the projection of $A$ to $G_1$, and for an $x \in A_1$ write

$$A(x) = \{y \in G_2 : (x, y) \in A\}.$$

Let

$$A' = \{(x, i) : x \in A_1, i \in \mathbb{Z},\ 0 \le i \le |A(x)| - 1\}$$

and

$$B' = \{(x, i) : x \in B_1, i \in \mathbb{Z},\ 0 \le i \le v_2\}.$$

We have $A', B' \subset G' = G_1 \times \mathbb{Z}$.

**Lemma 4.8.** *We have*

$$|A'| = |A|, \quad |A' + B'| \le |A + B|. \tag{4.6}$$

*Proof.* The equality is clear. To prove the inequality, write $S = A + B$, $S' = A' + B'$. With the obvious notation, we will show that

$$|S'(x)| \le |S(x)|$$

for each $x$. To this end observe that

$$S(x) = \bigcup_{x'+x''=x} (A(x') + B(x'')) = \bigcup_{x' \in x - B_1} A(x') + B_2,$$

hence

$$|S(x)| \geq \max_{x' \in x - B_1} |A(x') + B_2| \geq \max_{x' \in x - B_1} |A(x')| + v_2.$$

Similarly

$$S'(x) = \bigcup_{x'+x''=x} (A'(x') + B'(x'')) = \bigcup_{x' \in x - B_1} [0, |A(x')| + v_2 - 1],$$

and so

$$|S'(x)| = \max_{x' \in x - B_1} |A(x')| + v_2. \qquad \square$$

Now we continue the proof of the theorem. Decompose $A'$ into layers according to the value of the second component; write

$$A' = \bigcup_{i=0}^{k} L_i \times \{i\},$$

where $k = \max |A(x)|$, $L_i \subset G_1$. Write $|L_i| = m_i$. We have $L_0 \supset L_1 \supset \cdots \supset L_k$, consequently $m_0 \geq m_1 \geq \cdots \geq m_k$.

The set $S'$ is the union of the sets $(L_i + B_1) \times \{i + j\}$, $0 \leq i \leq v_2$. By the above inclusion it is sufficient to consider the $L_i$ with the smallest possible $i$, that is,

$$S' = (L_0 + B_1) \times \{0, 1, \ldots, v_2\} \cup \bigcup_{i=1}^{k} (L_i + B_1) \times \{i + v_2\}.$$

We obtain that

$$|S'| = v_2 |L_0 + B_1| + \sum_{i=0}^{k} |L_i + B_1|. \tag{4.7}$$

To estimate the summands we use the $d - 1$-dimensional impact of $B_1$, and we get

$$|L_i + B_1| \geq \left( m_i^{\frac{1}{d-1}} + v_1^{\frac{1}{d-1}} \right)^{d-1} \geq \frac{m_i}{m_0} \left( m_0^{\frac{1}{d-1}} + v_1^{\frac{1}{d-1}} \right)^{d-1};$$

the second inequality follows from $m_i \leq m_0$. By substituting this into (4.7) and recalling that $\sum m_i = m$ we obtain

$$|S| \geq \left( v_2 + \frac{m}{m_0} \right) \left( m_0^{\frac{1}{d-1}} + v_1^{\frac{1}{d-1}} \right)^{d-1}. \tag{4.8}$$

Consider the right side as a function of the real variable $m_0$. By differentiating we find that it assumes its minimum at

$$m_0 = v_1^{1/d} (m/v_2)^{1-1/d}.$$

(This minimum typically is not attained; this $m_0$ may be $< 1$ or $> m$, and it is generally not integer). Substituting this value of $m_0$ into (4.8) we obtain the desired bound

$$|S| \geq \left( m^{1/d} + (v_1 v_2)^{1/d} \right)^d. \qquad \square$$

**Problem 4.9.** Does equality always hold in Theorem 4.7?

I expect a negative answer.

**Problem 4.10.** Can Theorem 4.7 be extended to an inequality of the form

$$\mathrm{iv}_{d_1+d_2}(B_1 \times B_2) \geq \mathrm{iv}_{d_1}(B_1)\mathrm{iv}_{d_2}(B_2)?$$

*Proof of Theorem* 4.5. To prove $\geq$ we use induction on $d$. The case $d = 1$ is obvious, and Theorem 4.7 provides the inductive step.

This means that with the cube $B$ defined in (4.4) we have

$$|A + B| \geq \left(|A|^{1/d} + v^{1/d}\right)^d.$$

Equality can occur for infinitely many values of $|A|$, namely it holds whenever $A$ is also a cube of the form

$$A = \{(x_1, \ldots, x_d) \in \mathbb{Z}^d : 0 \leq x_i \leq kn_i - 1\}$$

with some integer $k$; we have $|A| = k^d v$, $|A + B| = (k + 1)^d v$. It may be difficult to describe $\xi_B(m)$ for values of $m$ which are not of the form $k^d v$. Possibly an argument like Gardner and Gronchi's for the simplex may work.

Observe that these special sets $A$ are not homothetic to $B$; in particular, $A = B$ may not yield a case of equality. □

As Theorem 4.3 shows, the impact volume can be $d!$ times smaller than cardinality. The example we have of this phenomenon, the long simplex, is, however, "barely" $d$-dimensional, and I expect that a better estimates hold for a "substantially" $d$-dimensional set.

**Definition 4.11.** The *thickness* $\vartheta(B)$ of a set $B \subset \mathbb{R}^d$ is the smallest integer $k$ with the property that there is a hyperplane $P$ of $\mathbb{R}^d$ and $x_1, \ldots, x_k \in \mathbb{R}^d$ such that $B \subset \bigcup_{i=1}^k P + x_j$.

**Conjecture 4.12.** For every $\varepsilon > 0$ and $d$ there is a $k$ such that for every $B \subset \mathbb{R}^d$ with $\vartheta(B) > k$ we have $\mathrm{iv}_d(B) > (1 - \varepsilon)|B|$.

This conjecture would yield a discrete Brunn–Minkowski inequality of the form

$$|A + B|^{1/d} \geq |A|^{1/d} + (1 - \varepsilon)|B|^{1/d}$$

assuming a bound on the thickness of $B$. Such an inequality is true at least in the special case $A = B$. This can be deduced from a result of Freiman ( [3], Lemma 2.12; see also Bilu [1]), which sounds as follows. If $A \subset \mathbb{R}^d$ and $|2A| < (2^d - \varepsilon)|A|$, then there is a hyperplane $P$ such that $|P \cap A| > \delta|A|$, with $\delta = \delta(d, \varepsilon) > 0$.

## 5. Meditation on the continuous case

Let $A, B$ be Borel sets in $\mathbb{R}^d$. The Brunn–Minkowski inequality (1.2) estimates $\mu(A+B)$ in a natural way, with equality if $A$ and $B$ are homothetic convex sets.

Like in the discrete case, we can define the *impact function* of the set $B$ by

$$\xi_B(a) = \inf\{\mu(A+B) : \mu(A) = a\}.$$

Thus (1.2) is equivalent to

$$\xi_B(a) \geq \left(a^{1/d} + \mu(B)^{1/d}\right)^d,$$

and this is the best possible estimate in terms of $\mu(B)$ only.

To measure the degree of nonconvexity we propose to use the measure of the convex hull beside the measure of the set. This is analogous to the hull volume, and it is sufficient to describe the asymptotic behaviour of $\xi$.

**Theorem 5.1** ([13], Theorem 1.). *For every bounded Borel set $B \subset \mathbb{R}^d$ of positive measure we have*

$$\lim_{a\to\infty} \xi_B(a)^{1/d} - a^{1/d} = \mu(\operatorname{conv} B)^{1/d}.$$

This is the continuous analogue of Theorem 3.6, and there is an analogue to the effective version Theorem 3.8 as well.

Note that by considering sets homothetic to conv $B$ we immediately obtain

$$\xi_B(a)^{1/d} \leq a^{1/d} + \mu(\operatorname{conv} B)^{1/d},$$

thus we need only to give a lower estimate. This is as follows.

**Theorem 5.2** ([13], Theorem 2.). *Let $\mu(B) = b$, $\mu(\operatorname{conv} B) = v$. We have*

$$\xi_B(a)^{1/d} \geq a^{1/d} + v^{1/d}\left(1 - c(v/b)^{1/2}(v/a)^{1/(2d)}\right)$$

$$\xi_B(a) \geq a + dv^{1/d}a^{1-1/d}\left(1 - c(v/b)^{1/2}(v/a)^{1/(2d)}\right)$$

*with a suitable positive constant $c$ depending on $d$.*

If $v > b$, we get a nontrivial improvement over the Brunn–Minkowski inequality for $a > a_0(b, v)$. It would be desirable to find an improvement also for small values of $a$, or, even more, to find the best estimate in terms of $\mu(B)$ and $\mu(\operatorname{conv} B)$.

The exact bound and the structure of the extremal set may be complicated. This is already so in the case $d = 1$, which was solved in [10]. Observe that in one dimension $\mu(\operatorname{conv} B)$ is the diameter of $B$.

**Theorem 5.3** ([10], Theorem 2). *Let $B \subset \mathbb{R}$, and write $\mu(B) = b$, $\mu(\operatorname{conv} B) = v$. If*

$$a \geq \frac{v(v-b)}{2b} + \frac{b\{v/b\}(1-\{v/b\})}{2}, \tag{5.1}$$

*then* $\xi_B(a) = a + v$. *If* (5.1) *does not hold, then let* $k$ *be the unique positive integer satisfying*

$$\frac{k(k-1)}{2} \le \frac{a}{b} < \frac{k(k+1)}{2}$$

*and define* $\delta$ *by*

$$\frac{a}{b} = \frac{k(k-1)}{2} + \delta k.$$

*We have*

$$\xi_B(a) \ge a + (k+\delta)b,$$

*and equality holds if* $B = [0, b] \cup \{v\}$.

A set $A$ such that $\xi_B(a) = \mu(A + B)$ for the above set $B$ is given by

$$A = [0, (k-1+\delta)b] \cup [v, v + (k-2+\delta)b] \cup \cdots \cup [(k-1)v, (k-1)v + \delta b].$$

A less exact, but simple and still quite good lower bound sounds as follows.

**Corollary 5.4** ([10], Theorem 1). *Let* $B \subset \mathbb{R}$, *and write* $\mu(B) = b$, $\mu(\operatorname{conv} B) = v$. *We have*

$$\xi_B(a) \ge \min\bigl(a + v, (\sqrt{a} + \sqrt{b/2})^2\bigr).$$

A comparison with the 2-dimensional Brunn–Minkowski inequality gives the following interpretation: initially a long one-dimensional set $B$ tries to behave as if it were a two-dimensional set of area $b/2$.

It can be observed that (5.4) is weaker than the obvious inequality

$$\mu(A + B) \ge \mu(A) + \mu(B) \tag{5.2}$$

for small $a$. For small values of $a$ Theorem 5.3 yields the following improvement of (5.2).

**Corollary 5.5** ([10], Corollary 3.1). *If* $a \le b$, *then we have*

$$\mu(A + B) \ge \min(2a + b, a + v).$$

*If* $b < a \le 3b$, *then we have*

$$\mu(A + B) \ge \min\left(\frac{3}{2}(a + b), a + v\right).$$

**Problem 5.6.** How large must $\mu(A+B)$ be if $\mu(A)$, $\mu(B)$, $\mu(\operatorname{conv} A)$ and $\mu(\operatorname{conv} B)$ are given?

What are the minima of $\mu(A + A)$ and $\mu(A - A)$ for fixed $\mu(A)$ and $\mu(\operatorname{conv} A)$?

The results above show that for $d = 1$ (like in the discrete case, but for less obvious reasons) the limit relation becomes an equality for $a > a_0$. Again, this is no longer the case for $d = 2$.

An example of a set $B \subset \mathbb{R}^2$ such that

$$\xi_B(a)^{1/2} < a^{1/2} + v^{1/2}$$

will hold for certain arbitrarily large values of $a$ is as follows.

Let $0 < c < 1$ and let $B$ consist of the square $[0, c] \times [0, c]$ and the points $(0, 1)$, $(1, 0)$ and $(1, 1)$. Hence $b = c^2$ and $v = 1$.

For an integer $n \geq 1$ put

$$A_n = [0, n] \times [0, n] \cup \bigcup_{j=0}^{n} [j, j + c] \times [n, n + c] \cup \bigcup_{j=0}^{n-1} [n, n + c] \times [j, j + c].$$

Thus $A_n$ consists of a square of side $n$ and $2n + 1$ small squares of side $c$, hence

$$\mu(A_n) = n^2 + (2n + 1)b.$$

We can easily see that $A_n + B = A_{n+1}$. Hence by considering the set $A = A_n$ we see that for a number $a$ of the form $a = n^2 + (2n + 1)b$ we have

$$\xi_B(a) \leq \mu(A_{n+1}) = (n + 1)^2 + (2n + 3)b < \left(\sqrt{a} + 1\right)^2.$$

A more detailed calculation leads to

$$\xi_B(a)^{1/2} \leq a^{1/2} + 1 - ca^{-1}$$

(for these special values of $a$).

If we tried to define an impact volume in the continuous case, we would recover the volume, at least for compact sets. Still, the above results and questions suggest that ordinary volume is not the best tool to understand additive properties. Perhaps one could try to modify the definition of impact volume by requiring $\mu(A) \geq \mu(B)$. So put

$$\mathrm{iv}_*(B) = \inf_{a \geq \mu(B)} \left(\xi_B(a)^{1/d} - a^{1/d}\right)^d.$$

**Problem 5.7.** Find a lower estimate for $\mathrm{iv}_*(B)$ in terms of $\mu(B)$ and $\mu(\mathrm{conv}\, B)$.

## 6. Back to one dimension

The results in the previous section, Theorem 5.3 and Corollaries 5.4 and 5.5 show that one can have nontrivial results in the seemingly uninteresting one-dimensional case. We now try to do the same, and will find bounds on $|A + B|$ using the cardinality and

hull volume of $B$. Observe that in one dimension the hull volume is the smallest $l$ such that $B$ is contained in an arithmetic progression $\{b, b+q, \dots, b+lq\}$: the *reduced diameter* of $B$.

It is possible to give bounds using nothing else than the hull volume.

**Theorem 6.1.** *Let $B$ be a one-dimensional set in a torsionfree commutative group,* $\mathrm{hv}\, B = v \geq 3$.

(a) *For*
$$m > \frac{(v-1)(v-2)}{2}$$
*we have $\xi_B(m) = m + v$.*

(b) *If*
$$\frac{(k-1)(k-2)}{2} < m \leq \frac{k(k-1)}{2}$$
*with some integer $2 \leq k < v$, then $\xi_B(m) \geq m + k$. Equality holds for the set $B = \{0, 1, v\} \subset \mathbb{Z}$.*

For $v \leq 2$ we have obviously $\xi_B(m) = m + v$ for all $m$ (such a set cannot be anything else than a $v+1$-term arithmetic progression).

This will be deduced from the following result, where the cardinality of $B$ is also taken into account.

**Theorem 6.2.** *Let $B$ be a one-dimensional set in a torsionfree commutative group,* $\mathrm{hv}\, B = v \geq 3$, $|B| = n$. *Define $w$ by*
$$w = \min_{d|v,\, d \leq n-2} d\left\lceil \frac{n-2}{d} \right\rceil. \tag{6.1}$$

*For every $m$ we have*
$$\xi_B(m) \geq m + \min\left(v, \frac{w}{2} + \min_{t \in \mathbb{N}} \left(\frac{m}{t} + \frac{tw}{2}\right)\right). \tag{6.2}$$

The minimum is attained either at the floor or at the ceiling of $\sqrt{2m/w}$. Unlike the previous theorem, typically we do not have examples of equality, and the extremal value and the structure of extremal sets is probably complicated. Also the value of $w$ depends on divisibility properties of $v$ and $n$. After the proof we give a less exact but simpler corollary.

*Proof.* By Lemma 3.1 we may assume that $B \subset \mathbb{Z}$, its smallest element is 0 and it generates $\mathbb{Z}$; then its largest element is just $v$.

**Lemma 6.3.** *Let $B'$ be the set of residues of elements of $B$ modulo $v$. For every nonempty $X \subset \mathbb{Z}_v$ we have*
$$|X + B'| \geq \min(|X| + w, v). \tag{6.3}$$

*Proof.* By Kneser's theorem we have

$$|X + B'| \geq |X + H| + |B' + H| - |H|$$

with some subgroup $H$ of the additive group $\mathbb{Z}_v$. Write $|H| = d$; clearly $d|v$. If $d = v$, we have $|X + H| = v$ and we are ready. Assume $d < v$. $B'$ contains 0 and it generates $\mathbb{Z}_v$, hence it cannot be contained in $H$ so we have $|B' + H| \geq 2|H| = 2d$. This gives the desired bound if $d > n - 2$. Assume $d \leq n - 2$. Since $|B' + H|$ is a multiple of $d$ and it is at least $|B'| = n - 1$, we obtain

$$|B' + H| \geq d \left\lceil \frac{n-1}{d} \right\rceil = d\left(1 + \left\lceil \frac{n-2}{d} \right\rceil\right) \geq d + w. \qquad \square$$

We resume the proof of Theorem 6.2. Take a set $A \subset \mathbb{Z}$, $|A| = m$. We are going to estimate $|A + B|$ from below.

For $j \in \mathbb{Z}_v$ let $u(j)$ be the number of integers $a \in A$, $a \equiv j \pmod{v}$ and let $U(j)$ be the corresponding number for the sumset $A + B$. We have

$$U(j) \geq u(j) + 1 \tag{6.4}$$

whenever $U(j) > 0$; this follows by adding the numbers $0, v$ to each element of $A$ in this residue class if $u(j) > 0$, and holds obviously for $u(j) = 0$. We also have

$$U(j) \geq u(j - b) \tag{6.5}$$

for every $b \in B'$. Write

$$r(k) = \{j : u(j) \geq k\},$$

$$R(k) = \{j : U(j) \geq k\}.$$

Inequality (6.4) implies

$$R(k) \supset r(k - 1) \quad (k \geq 2), \tag{6.6}$$

and inequality (6.5) implies

$$R(k) \supset r(k) + B' \quad (k \geq 1). \tag{6.7}$$

*First case.* $U(j) > 0$ for all $j$. In this case by summing (6.4) we get

$$|A + B| = \sum U(j) \geq v + \sum u(j) = |A| + v.$$

*Second case.* There is a $j$ with $U(j) = 0$. Then we have $|R(k)| < v$ for every $k > 0$. An application of Lemma 6.3 to the sets $r(k)$ yields, in view of (6.7),

$$|R(k)| \geq |r(k)| + w \tag{6.8}$$

as long as $r(k) \neq \emptyset$. Let $t$ be the largest integer with $r(t) \neq \emptyset$. We have (6.8) for $1 \leq k \leq t$, and (6.6) yields

$$|R(k)| \geq |r(k-1)| \tag{6.9}$$

for all $k \geq 2$. Consequently for $1 \leq k \leq t+1$ we have

$$|R(k)| \geq \frac{k-1}{t}|r(k-1)| + \left(1 - \frac{k-1}{t}\right)(|r(k)| + w). \tag{6.10}$$

Indeed, for $k = 1$ (6.10) is identical with (6.8), for $k = t+1$ it is identical with (6.9) and for $2 \leq k \leq t$ it is a linear combination of the two.

By summing (6.10) we obtain

$$\begin{aligned} |A+B| &= \sum_{k\geq 1} |R(k)| \geq \sum_{k=1}^{t+1} |R(k)| \geq \frac{t+1}{2}w + \left(1 + \frac{1}{t}\right)\sum_{k=1}^{t} |r(k)| \\ &= \frac{t+1}{2}w + \left(1 + \frac{1}{t}\right)|A|, \end{aligned}$$

as claimed in (6.2). □

**Corollary 6.4.** *With the assumptions and notations of Theorem* 6.2 *we have*

$$\xi_B(m) \geq \min\left(m + v, \left(\sqrt{m} + \sqrt{w/2}\right)^2\right). \tag{6.11}$$

*Proof.* This follows from (6.2) and the inequality of arithmetic and geometric means. □

This can be interpreted as that the set tries to imitate a two-dimensional set of area $w/2$.

*Proof of Theorem* 6.1. Parts (a)–(b) of the theorem can be reformulated as follows: if $\xi_B(n) \leq m + k$ with some $k < v$, then $m \leq k(k-1)/2$. Theorem 6.2 yields (using only that $w \geq 1$) the existence of a positive integer $t$ such that

$$\frac{m}{t} + \frac{t+1}{2} \leq k,$$

hence

$$m \leq kt + \frac{t(t+1)}{2}.$$

The right side, as a function of $t$, is increasing up to $k - 1/2$ and decreasing afterwards; the minimal values at integers are assumed at $t = k-1$ and $k$, and both are equal to $k(k-1)/2$.

To show the case of equality in case (b), write $m = k(k-1)/2 - l$ with $0 \leq l \leq k-2$. The set $A$ will contain the integers in the intervals $[iv, iv + k - 3 - i]$ for $0 \leq i \leq l-1$ and $[iv, iv + k - 2 - i]$ for $l \leq i \leq k-2$. □

We illustrate the strength of Theorem 6.2 by deducing from it the two-dimensional estimate

$$\xi_L(m) > \left(\sqrt{m} + \sqrt{(n-2)/2}\right)^2$$

for the long triangle $L = L_{2n}$. Indeed, a suitable linear mapping maps this set $L$ onto the set $B = \{0, 1, \dots, n-2, v\}$ with arbitrary $v$. If we choose $v$ to be prime, then in (6.1) we have $w = n-2$, and if $v$ is so large that $m + l > \left(\sqrt{m} + \sqrt{w}\right)^2$, then from Corollary 6.4 we obtain

$$\xi_L(m) \geq \xi_B(m) \geq \left(\sqrt{m} + \sqrt{w/2}\right)^2.$$

This is essentially the two-dimensional case of Theorem 2.6 of Gardner and Gronchi.

On the other hand, for small values of $m$ this inequality is weak, can even be worse than the obvious bound $|A + B| \geq |A| + |B| - 1$. There are results that are especially suited to the study of small values; we quote two of them. In both let $A, B \subset \mathbb{Z}$, $A = \{a_1, \dots, a_m\}$, $B = \{b_1, \dots, b_n\}$ with $0 = a_1 < \cdots < a_m = u$, $0 = b_1 < \cdots < b_n = v$.

**Theorem 6.5** (Freiman [2]). *If* $\gcd(a_1, \dots, a_m, b_1, \dots, b_n) = 1$ *and* $u \leq v$, *then*

$$|A + B| \geq \min(m + v, m + n + \min(m, n) - 3).$$

This bears a remarkable similarity to the two-dimensional case of Theorem 1.2 (and it can be deduced like Theorem 2.6)

**Theorem 6.6** (Lev and Smelianski [7]). *If* $\gcd(b_1, \dots, b_n) = 1$ *and* $u \leq v$, *then*

$$|A + B| \geq \min(m + v, n + 2m - \delta),$$

*where* $\delta = 3$ *if* $u = v$ *and* $\delta = 2$ *if* $u < v$.

Observe that the above theorems cannot be directly compared to ours because of the somewhat different structure of the assumptions.

**Problem 6.7.** Find a common generalization of Theorems 6.2 and 6.6.

## References


[1] Bilu, Y., Structure of sets with small sumset. Structure theory of set addition. *Astérisque* **258** (1999), 77–108.

[2] Freiman, G., Inverse problems of additive number theory. VI. On the addition of finite sets. III. *Izv. Vyss. Ucebn. Zaved. Matematika* **3** (28) (1962), 151–157 (in Russian).

[3] —, *Foundations of a structural theory of set addition*. Transl. Math. Monogr. 37, Amer. Math. Soc., Providence, RI, 1973.

[4] Gardner, R. J., and Gronchi, P., A Brunn-Minkowski inequality for the integer lattice. *Trans. Amer. Math. Soc.* **353** (2001), 3995–4024.

[5] Khovanskii, A. G., Newton polyhedron, Hilbert polynomial, and sums of finite sets. *Funct. Anal. Appl.* **26** (1992), 276–281.

[6] —, Sums of finite sets, orbits of commutative semigroups, and hilbert functions. *Funct. Anal. Appl.* **29** (1995), 102–112.

[7] Lev, V. F., and Smeliansky, P., On addition of two distinct sets of integers. *Acta Arith.* **70** (1995), 85–91.

[8] Nathanson, M. B., and Ruzsa, I. Z., Polynomial growth of sumsets in abelian semigroups. *J. Théor. Nombres Bordeaux* **14** (2002), 553–560.

[9] Plünnecke, H., *Eigenschaften und Abschätzungen von Wirkungsfunktionen*. Gesellschaft für Mathematik und Darenverarbeitung, Bonn, 1969.

[10] Ruzsa, I. Z., Diameter of sets and measure of sumsets. *Monatsh. Math.* **112** (1991), 323–328.

[11] —, Sum of sets in several dimensions. *Combinatorica* **14** (1994), 485–490.

[12] —, Sets of sums and commutative graphs. *Studia Sci. Math. Hungar.* **30** (1–2) (1995), 127–148.

[13] —, The Brunn-Minkowski inequality and nonconvex sets. *Geom. Dedicata* **67** (1997), 337–348.

Alfréd Rényi Institute of Mathematics, 1364 Budapest, Pf. 127, Hungary
E-mail: ruzsa@renyi.hu

# Geometric bistellar flips: the setting, the context and a construction

Francisco Santos *

**Abstract.** We give a self-contained introduction to the theory of secondary polytopes and geometric bistellar flips in triangulations of polytopes and point sets, as well as a review of some of the known results and connections to algebraic geometry, topological combinatorics, and other areas.

As a new result, we announce the construction of a point set in general position with a disconnected space of triangulations. This shows, for the first time, that the poset of strict polyhedral subdivisions of a point set is not always connected.

**Mathematics Subject Classification (2000).** Primary 52B11; Secondary 52B20.

**Keywords.** Triangulation, point configuration, bistellar flip, polyhedral subdivision, disconnected flip-graph.

## Introduction

Geometric bistellar flips are "elementary moves", that is, minimal changes, between triangulations of a point set in affine space $\mathbb{R}^d$. In their present form they were introduced around 1990 by Gel'fand, Kapranov and Zelevinskii during their study of discriminants and resultants for sparse polynomials [28], [29]. Not surprisingly, then, these bistellar flips have several connections to algebraic geometry. For example, the author's previous constructions of point sets with a disconnected graph of triangulations in dimensions five and six [64], [67] imply that certain algebraic schemes considered in the literature [4], [13], [33], [57], including the so-called toric Hilbert scheme, are sometimes not connected.

Triangulations of point sets play also an obvious role in applied areas such as computational geometry or computer aided geometric design, where a region of the plane or 3-space is triangulated in order to approximate a surface, answer proximity or visibility questions, etc. See, for example, the survey articles [8], [10], or [25]. In these fields, flips between triangulations have also been considered since long [40]. Among other things, they are used as the basic step to compute an optimal triangulation of a point set incrementally, that is, adding the points one by one. This *incremental flipping algorithm* is the one usually preferred for, for example, computing the Delaunay

*Partially supported by the Spanish Ministry of Education and Science, grant number MTM2005-08618-C02-02.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

triangulation, as "the most intuitive and easy to implement" [8], and yet as efficient as any other.

In both the applied and the theoretical framework, the situation is the same: a fixed set of points $\mathcal{A} \subset \mathbb{R}^d$ is given to us (the "sites" for a Delaunay triangulation computation, the test points for a surface reconstruction, or a set of monomials, represented as points in $\mathbb{Z}^d$, in the algebro-geometric context) and we need to either explore the collection of all possible triangulations of this set $\mathcal{A}$ or search for a particular one that satisfies certain optimality properties. Geometric bistellar flips are the natural way to do this. For this reason, it was considered one of the main open questions in polytope theory ten years ago whether point sets exist with triangulations that cannot be connected via these flips [80]. As we have mentioned above, this question was answered positively by the author of this paper, starting in dimension five. The question is still open in dimensions three and four.

This paper intends to be an introduction to this topic, organized in three parts.

The first section is a self-contained introduction to the theory of geometric bistellar flips and *secondary polytopes* in triangulations of point sets, aimed at the non-expert. The results in it are certainly not new (most come from the original work of Gel'fand, Kapranov and Zelevinskii mentioned above) but the author wants to think that this section has some expository novelty; several examples that illustrate the theory are given, and our introduction of geometric bistellar flips first as certain polyhedral subdivisions and only afterwards as transformations between triangulations is designed to show that the definition is as natural as can be. This section finishes with an account of the state-of-the-art regarding knowledge of the graph of flips for sets with "few" points or "small" dimension, with an emphasis on the differences between dimensions two and three.

The second section develops in more detail the two contexts in which we have mentioned that flips are interesting (*computational geometry* and *algebraic geometry*) together with other two, that we call "*combinatorial topology*" and "*topological combinatorics*". Combinatorial topology refers to the study of topological manifolds via triangulations of them. Bistellar flips have been proposed as a tool for manifold recognition [18], [46], and triangulations of the 3-sphere without bistellar flips other than "insertion of new vertices" are known [24]. Topological combinatorics refers to topological methods in combinatorics, particularly to the topology of *partially ordered sets* (posets) via their order complexes. The graph of triangulations of a point set $\mathcal{A}$ consists of the first two levels in the poset of polyhedral subdivisions of $\mathcal{A}$, which in turn is just an instance of several similar posets studied in combinatorics with motivations and applications ranging from oriented matroid theory to bundle theories in differential geometry.

The third section announces for the first time the construction of a point set *in general position* whose graph of triangulations is not connected. The details of the proof appear in [68]. The point set is also the smallest one known so far to have a disconnected graph of flips.

**Theorem.** *There is a set of* 17 *points in general position in* $\mathbb{R}^6$ *whose graph of triangulations is not connected.*

As usual in geometric combinatorics, a finite point set $\mathcal{A} \subset \mathbb{R}^d$ is said to be *in general position* if no $d+2$ of the points lie in an affine hyperplane. Equivalently, if none of the $\binom{|\mathcal{A}|}{d+1}$ determinants defined by the point set vanish. Point sets in general position form an open dense subset in the space $\mathbb{R}^{n\times d}$ of sets of dimension $d$ with $n$ elements. That is to say, "random point sets" are in general position. Point sets that are not in general position are said to be *in special position*.

The connectivity question has received special attention in general position even before disconnected examples in special position were found. For example, Challenge 3 in [80] and Problem 28 in [50] specifically ask whether disconnected graphs of flips exist for point sets in special position (the latter asks this only for dimension 3). Although it was clear (at least to the author of this paper) from the previous examples of disconnected graphs of flips that examples in general position should also exist, modifying those particular examples to general position and proving that their flip-graphs are still not connected is not an easy task for quite intrinsic reasons: the proofs of non-connectednes in [64], [67] are based on the fact that the point sets considered there are cartesian products of lower dimensional ones.

In our opinion, an example of a disconnected graph of flips in general position is interesting for the following three reasons:

1. The definition of flip that is most common in computational geometry coincides with ours (which is the standard one in algebraic geometry and polytope combinatorics) only for point sets in general position. In special position, the computational geometric definition is far more restrictive and, in particular, taking it makes disconnected graphs of flips in special position be "no surprise". For example, Edelsbrunner [25] says that the flip-graph among the (three) triangulations of a regular octahedron is not connected; see Section 2.1.
2. Leaving aside the question of definition, in engineering applications the coordinates of points are usually approximate and there is no loss in perturbing them into general position. That is, the general position case is sometimes the only case.
3. Even in a purely theoretical framework, point sets in general position have somehow simpler properties than those in special position. If a point set $\mathcal{A}$ in special position has a non-connected graph of flips then automatically some subset of $\mathcal{A}$ (perhaps $\mathcal{A}$ itself) has a disconnected poset of subdivisions. This poset is sometimes called the *Baues poset* of $\mathcal{A}$ and its study is (part of) the so-called generalized Baues problem. See Section 2.3, or [61] for more precise information on this. In particular, the present example is the first one (proven) to have a disconnected Baues poset.

**Corollary.** *There is a set of at most* 17 *points in* $\mathbb{R}^6$ *whose poset of proper polyhedral subdivisions is not connected.*

# 1. The setting

## 1.1. Triangulations. Regular triangulations and subdivisions

**Triangulations and polyhedral subdivisions.** A (convex) *polytope* $P$ is the convex hull of a finite set of points in the affine space $\mathbb{R}^d$. A *face* of $P$ is its intersection with any hyperplane that does not cross the relative interior of $P$. (Here, the *relative interior* of $S \subseteq \mathbb{R}^d$ is the interior of $S$ regarded as a subset of its affine span). We remind the reader that the faces of dimensions $0$, $1$, $d-2$ and $d-1$ of a $d$-polytope are called vertices, edges, ridges and facets, respectively. Vertices of $P$ form the minimal $S$ such that $P = \text{conv}(S)$.

A $k$-simplex is a polytope whose vertices (necessarily $k+1$) are affinely independent. It has $\binom{k+1}{i+1}$ faces of each dimension $i = 0, \ldots, k$, which are all simplices.

**Definition 1.1.** Let $\mathcal{A}$ be a finite point set in $\mathbb{R}^d$. A *triangulation* of $\mathcal{A}$ is any collection $T$ of affinely spanning and affinely independent subsets of $\mathcal{A}$ with the following properties:

1. if $\sigma$ and $\sigma'$ are in $T$, then $\text{conv}(\sigma) \cap \text{conv}(\sigma')$ is a face of both $\text{conv}(\sigma)$ and $\text{conv}(\sigma')$. That is, $T$ induces a geometric simplicial complex in $\mathbb{R}^k$;
2. $\bigcup_{\sigma \in T} \text{conv}(\sigma) = \text{conv}(\mathcal{A})$. That is, $T$ covers the convex hull of $\mathcal{A}$.

Note that our definition allows for some points of $\mathcal{A}$ not to be used at all in a particular triangulation. Extremal points (vertices of $\text{conv}(\mathcal{A})$) are used in every triangulation. The elements of a triangulation $T$ are called *cells*.

We can define *polyhedral subdivisions* of $\mathcal{A}$ by removing the requirement of the sets $\sigma$ to be affinely independent in Definition 1.1. Since a general subset $\sigma$ of $\mathcal{A}$ may contain points which are not vertices of $\text{conv}(\sigma)$, now the fact that the elements of a subdivision are subsets of $\mathcal{A}$ rather than "subpolytopes" is not just a formality: points which are not vertices of any "cell" in the subdivision may still be considered "used" as elements of some cells. In order to get a nicer concept of polyhedral subdivision, we also modify part 1 in Definition 1.1, adding the following (redundant for affinely independent sets) condition:

$$\text{conv}(\sigma \cap \sigma') \cap \sigma = \text{conv}(\sigma \cap \sigma') \cap \sigma' \quad \text{for all } \sigma, \sigma' \in T.$$

That is, if $\mathcal{A}$ contains some point in the common face $\text{conv}(\sigma \cap \sigma')$ of $\text{conv}(\sigma)$ and $\text{conv}(\sigma')$ but not a vertex of it, that point is either in both or in none of $\sigma$ and $\sigma'$.

Polyhedral subdivisions of $\mathcal{A}$ form a *partially ordered set* (or *poset*) with respect to the following refinement relation:

$$S \text{ refines } S' \ :\Leftrightarrow \ \text{for all } \sigma' \in S' \text{ there exists } \sigma \in S \text{ such that } \sigma \subseteq \sigma'.$$

Triangulations are, of course, the minimal elements in this poset. The poset has a unique maximal element, namely the *trivial suvdivision* $\{\mathcal{A}\}$.

**Example 1.2.** Let $\mathcal{A}$ be the following set of five points $a_1, \ldots, a_5$ in the plane. We take the convention that points are displayed as columns in a matrix, and that an extra homogenization coordinate (the row of 1's in the following matrix) is added so that linear algebra, rather than affine geometry, can be used for computations:

$$\mathcal{A} = \begin{array}{c} \begin{array}{ccccc} a_1 & a_2 & a_3 & a_4 & a_5 \end{array} \\ \begin{pmatrix} 0 & 3 & 0 & 3 & 1 \\ 0 & 0 & 3 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array} \tag{1}$$

The following are the nine polyhedral subdivisions of $\mathcal{A}$. Arrows represent the refinement relation, pointing from the coarser to the finer subdivision. For clarity, we write "125" meaning $\{a_1, a_2, a_5\}$, and so on. Figure 1 shows pictures of the subdivisions. In the corners are the four triangulations of $\mathcal{A}$ and in the middle is the trivial subdivision.

$$\begin{array}{ccccc} \{125, 135, 235, 234\} & \leftarrow & \{1235, 234\} & \rightarrow & \{135, 234\} \\ \uparrow & & \uparrow & & \uparrow \\ \{125, 135, 2345\} & \leftarrow & \{12345\} & \rightarrow & \{1234\} \\ \downarrow & & \downarrow & & \downarrow \\ \{125, 135, 245, 345\} & \leftarrow & \{1245, 1345\} & \rightarrow & \{124, 134\} \end{array}$$

The last two columns of subdivisions geometrically induce the same decomposition

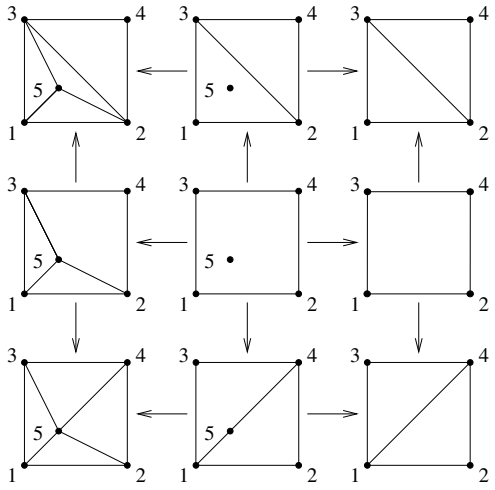



Figure 1. The nine polyhedral subdivisions of a certain point set.

of conv($\mathcal{A}$) into subpolygons. Still, we consider them different subdivisions since the middle column "uses" the interior point 5 while the right column does not.

**Regular subdivisions.** Let a point set $\mathcal{A}$ be given, and choose a function $w\colon \mathcal{A} \to \mathbb{R}$ to lift $\mathcal{A}$ to $\mathbb{R}^{d+1}$ as the point set

$$\mathcal{A}_w := \{(a, w(a) : a \in \mathcal{A}\}.$$

A *lower facet* of $\operatorname{conv}(\mathcal{A}_w)$ is a facet whose supporting hyperplane lies below the interior of $\operatorname{conv}(\mathcal{A}_w)$. The following is a polyhedral subdivision of $\mathcal{A}$, where $\pi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ is the projection that forgets the last coordinate:

$$T_w := \{\pi(F \cap \mathcal{A}_w) : F \text{ is a lower facet of } \operatorname{conv}(\mathcal{A}_w)\}.$$

Geometrically, we are projecting down onto $\mathcal{A}$ the lower envelope of $\mathcal{A}_w$, keeping track of points that lie in the lower boundary even if they are not vertices of a facet.

**Definition 1.3.** The polyhedral subdivisions and triangulations that can be obtained in this way are called *regular*.

If $w$ is sufficiently generic then $T_w$ is clearly a triangulation. Regular triangulations are particularly simple and yet quite versatile. They appear in different contexts under different names such as *coherent* [29], *convex* [36], [77], *Gale* [49], or *generalized* (*or, weighted*) *Delaunay* [25] triangulations. The latter refers to the fact that the Delaunay triangulation of $\mathcal{A}$, probably the most used triangulation in applications, is the regular triangulation obtained with $w(a) = \|a\|^2$, where $\|\cdot\|$ is the euclidean norm.

**Example 1.4.** Let

$$\mathcal{A} = \begin{array}{c} \begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{matrix} \\ \begin{pmatrix} 4 & 0 & 0 & 2 & 1 & 1 \\ 0 & 4 & 0 & 1 & 2 & 1 \\ 0 & 0 & 4 & 1 & 1 & 2 \end{pmatrix} \end{array}.$$

This is a configuration of six points in the affine plane with equation $x_1 + x_2 + x_3 = 4$ in $\mathbb{R}^3$. Since the matrix is already homogeneous (meaning precisely that columns lie in an affine hyperplane) we do not need the extra homogenization row. The configuration consists of two parallel equilateral triangles, one inside the other. We leave it to the reader to check that the following are two non-regular triangulations (see Figure 2):

$$T_1 := \{124, 235, 136, 245, 356, 146, 456\},$$
$$T_2 := \{125, 236, 134, 145, 256, 346, 456\}.$$

This example is the smallest possible, since 1-dimensional point configurations and point configurations with at most $d + 3$ points in any dimension $d$ only have regular triangulations. The former is easy to prove and the latter was first shown in [44]. The earliest appearance of these two non-regular triangulations that we know of is in [20], although they are closely related to Schönhardt's classical example of a non-convex 3-polytope that cannot be triangulated [69].[1]

[1] We describe Schönhardt's polyhedron and its relation to this example in Example 1.21.

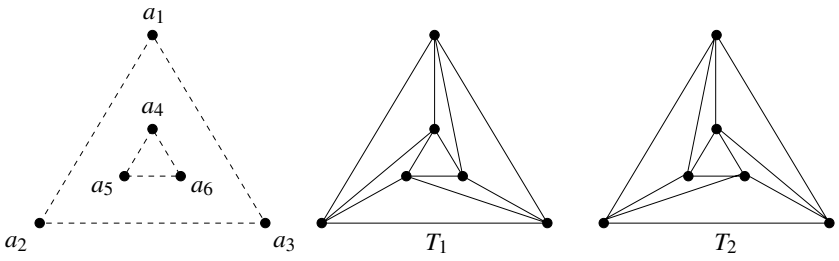



Figure 2. A point configuration with two non-regular triangulations.

**Remark 1.5.** Suppose that two point sets $\mathcal{A} = \{a_1, \ldots, a_n\}$ and $\mathcal{B} = \{b_1, \ldots, b_n\}$ have the same *oriented matroid* [17], or *order type*. This means that for every subset $I \subset \{1, \ldots, n\}$ of labels, the determinants of the point sets $\{a_i : i \in I\}$ and $\{b_i : i \in I\}$ have the same sign.[2] It is an easy exercise to check that then $\mathcal{A}$ and $\mathcal{B}$ have *the same* triangulations and subdivisions.[3] However, they do not necessarily have the same *regular* subdivisions. For example, the points of example 1.4 are in general position and, hence, their oriented matroid does not change by a small perturbation of coordinates. But any sufficiently generic perturbation makes one of the two non-regular triangulations $T_1$ and $T_2$ become regular.

Still, the following is true [65]: the *existence* of non-regular triangulations of $\mathcal{A}$ depends only on the oriented matroid of $\mathcal{A}$.

**The secondary polytope.** Let $L_{\mathcal{A}}$ denote the space of all lifting functions $w : \mathcal{A} \to \mathbb{R}$ on a certain point set $\mathcal{A} \subset \mathbb{R}^d$ with $n$ elements. In principle $L_{\mathcal{A}}$ is isomorphic to $\mathbb{R}^n$ in an obvious way; but we mod-out functions that lift all of $\mathcal{A}$ to a hyperplane, because adding one of them to a given lifting function $w$ does not (combinatorially) change the lower envelope of $\mathcal{A}_w$. We call these particular lifting functions *affine*. They form a linear subspace of dimension $d+1$ of $\mathbb{R}^n$. Hence, after we mod-out affine functions we have $L_{\mathcal{A}} \cong \mathbb{R}^{n-d-1}$.

For a given polyhedral subdivision $T$ of $\mathcal{A}$, the subset of $L_{\mathcal{A}}$ consisting of functions $w$ that produce $T = T_w$, is a (relatively open) polyhedral cone; that is, it is defined by a finite set of linear homogeneous equalities and strict inequalities. Equalities appear only if $T$ is not a triangulation and express the fact that if $\sigma \in T$ is not affinely independent then $w$ must lift all $\sigma$ to lie in a hyperplane. Inequalities express the fact that for each $\sigma \in T$ and point $a \in \mathcal{A} \setminus \sigma$, $a$ is lifted above the hyperplane spanned by the lifting of $\sigma$.

The polyhedral cones obtained for different choices of $T$ are glued together forming a polyhedral fan, that is, a "cone over a polyhedral complex", called the *secondary fan* of $\mathcal{A}$. The prototypical example of a fan is the normal fan of a polytope, whose

[2] Observe that the bijection between $\mathcal{A}$ and $\mathcal{B}$ implicit by the labels is part of the definition.

[3] More precisely, the implicit bijection between $\mathcal{A}$ and $\mathcal{B}$ induces a bijection between their polyhedral subdivisions.

cones are the exterior normal cones of different faces of $P$. A seminal result in the theory of triangulations of polytopes is that the secondary fan is actually polytopal; that is, it is the normal fan of a certain polytope:

**Theorem 1.6** (Gel'fand–Kapranov–Zelevinskii [28], [29]). *For every point set $\mathcal{A}$ of $n$ points affinely spanning $\mathbb{R}^d$ there is a polytope $\Sigma(\mathcal{A})$ in $L_{\mathcal{A}} \cong \mathbb{R}^{n-d-1}$ whose normal fan is the secondary fan of $\mathcal{A}$.*

In particular, the poset of regular subdivisions of $\mathcal{A}$ is isomorphic to the poset of faces of $\Sigma(\mathcal{A})$. Vertices correspond to regular triangulations and $\Sigma(\mathcal{A})$ itself (which is, by convention, considered a face) corresponds to the trivial subdivision. The polytope $\Sigma(\mathcal{A})$ is called the *secondary polytope of* $\mathcal{A}$.

There are two standard ways to construct the secondary polytope $\Sigma(\mathcal{A})$ of a point set $\mathcal{A}$.[4] The original one, by Gel'fand, Kapranov and Zelevinskii [28], [29] gives, for each regular triangulation $T$ of $\mathcal{A}$, coordinates of the corresponding vertex $v_T$ of $\Sigma(\mathcal{A})$ in terms of the volumes of simplices incident in $T$ to each point of $\mathcal{A}$.

The second one, by Billera and Sturmfels [14], describes the whole polytope $\sigma(\mathcal{A})$ as the Minkowski integral of the fibers of the affine projection $\pi : \Delta_{\mathcal{A}} \to \text{conv}(\mathcal{A})$, where $\Delta_{\mathcal{A}}$ is a simplex with $|\mathcal{A}|$ vertices (hence, of dimension $|\mathcal{A}| - 1$) and $\pi$ bijects the vertices of $\Delta_{\mathcal{A}}$ to $\mathcal{A}$ (see Theorem 2.8).

**Example 1.7** (Example 1.2 continued). Figure 3 shows the secondary fan of the five points. To mod-out affine functions we have taken $w(a_1) = w(a_2) = w(a_3) = 0$, and the horizontal and vertical coordinates in the figure give the values of $w(a_4)$ and $w(a_5)$, respectively. The triangulation corresponding to each two-dimensional cone is displayed. In this example all nine polyhedral subdivisions are regular (in agreement

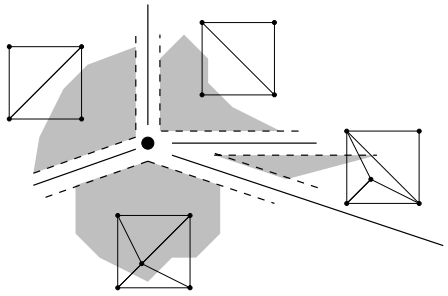


Figure 3. The secondary fan of Example 1.2.

with the result of [44] mentioned in Example 1.4) and the secondary polytope is a quadrilateral.

[4] Polytopality of a fan is equivalent to the feasibility of a certain system of linear equalities and strict inequalities. But here we mean more direct and intrinsic constructions of the secondary polytope.

**Example 1.8** (Example 1.4 continued). The secondary polytope of this point set is 3-dimensional, and contains a hexagonal face corresponding to the regular subdivision

$$\{1245, 2356, 1346, 456\}.$$

This regular subdivision can be refined to a triangulation in eight ways, by independently inserting a diagonal in the quadrilaterals 1245, 2356 and 1346. Six of these triangulations are regular, and correspond to the vertices of the hexagon. The other two, $T_1$ and $T_2$, are non-regular and they "lie" in the center of the hexagon.

We have mentioned that if the point set is perturbed slightly then one of the triangulations becomes regular. What happens in the secondary polytope is the following: the perturbation "inflates" the hexagon so that the eight points (the vertices of the hexagon and the two interior points representing $T_1$ and $T_2$) become, combinatorially, the vertices of a cube. The points corresponding to $T_1$ and $T_2$ move in opposite directions, one of them going to the interior of the secondary polytope and the other becoming a new vertex of it. The hexagonal face gets refined into three quadrilaterals. Of course, the vertices of the hexagon also move in the process, and are no longer coplanar.

**Example 1.9** (The convex $n$-gon and the associahedron). All triangulations of a convex $n$-gon are regular and their number is the $n-2$nd Catalan number

$$C_{n-2} := \frac{1}{n-1}\binom{2n-4}{n-2}.$$

The corresponding secondary polytope is called the *associahedron*. The name comes from the fact that there is a bijection between triangulations of the $n$-gon and the ways to put the $n-2$ parentheses in an associative product of $n-1$ factors.

The associahedron is a classical object in combinatorics, first studied[5] by Stasheff and Tamari [76], [72]. It was shown to be polytopal by Haiman (unpublished) and Lee [43]. That its diameter equals $2n-10$ "for every sufficiently big $n$"[6] was shown by Sleator, Tarjan and Thurston [71], with motivations coming from theoretical computer science and tools from hyperbolic geometry.

**Remark 1.10.** In Sections 2.3 and 2.4 we will mention triangulations of a set of *vectors* rather than *points*. They are defined exactly as triangulations of point sets, just changing the word "affinely" to "linearly" and the operator "conv" to "pos" ("positive span") in Definition 1.1. Put differently, a triangulation of a vector set $\mathcal{A} \subset \mathbb{R}^{d+1}$ is a simplicial fan covering $\mathrm{pos}(\mathcal{A})$ and whose rays are in the positive directions of (not necessarily all) the elements of $\mathcal{A}$. Equivalently, and perhaps closer to readers familiar with classical geometry, we can, without loss of generality, normalize all

[5] As a combinatorial cell complex, without an explicit polytopal realization.

[6] Sleator et al. do not say "how big" is "sufficiently big" in their statement, but conjecture that $n \geq 13$ is enough. We consider this an interesting and somehow shameful open question.

vectors of $\mathcal{A}$ to lie in the unit sphere $S^d$. Then, triangulations of $\mathcal{A}$ are the geodesic triangulations, with vertices contained in $\mathcal{A}$, of the spherical convex hull of $\mathcal{A}$.

The existence and properties of regular subdivisions and secondary fans (and of the bistellar flips introduced in the next section) generalize almost without change to vector configurations.[7]

### 1.2. Geometric bistellar flips

**Flips as polyhedral subdivisions.** In order to introduce the notion of local move (flip) between triangulations of $\mathcal{A}$, we use the secondary fan as a guiding light: whatever our definition is, restricted to regular triangulations a flip should correspond to crossing a "wall" between two full-dimensional cones in the secondary fan; that is, a flip between two regular triangulations $T_1$ and $T_2$ can be regarded as certain regular subdivision $T_0$ with the property that its only two regular refinements are precisely $T_1$ and $T_2$. Some thought will convince the reader that the necessary and sufficient condition for a lifting function $w\colon \mathcal{A} \to \mathbb{R}$ to produce a $T_w$ with this property is that *there is a unique minimal affinely dependent subset in $\mathcal{A}$ whose lifting is contained in some lower facet of the lifted point set $\mathcal{A}_w$*. This leads to the following simple, although perhaps not very practical, definition.

**Definition 1.11.** Let $T$ be a (not-necessarily regular) subdivision of a point set $\mathcal{A}$. We say that $T$ *is a flip* if there is a unique affinely dependent subset $C \in \mathcal{A}$ contained in some cell of $T$.

**Lemma 1.12.** *If $T$ is a flip, then there are exactly two proper refinements of $T$, which are both triangulations.*

*Proof.* Let $T_1$ be a refinement of $T$. Let $C$ be the unique affinely dependent subset of $\mathcal{A}$ contained in some cell of $T$. Each cell of $T$ containing $C$ gets refined in $T_1$, while each cell not containing $C$ is also a cell in $T_1$.

The statement then follows from the understanding of the combinatorics of point sets with a unique affinely dependent subset $C$. Let $S$ be such a point set. Each point in $S \setminus C$ is affinely independent of the rest, so $S$ is an "iterated cone" over $C$. In particular, there is a face $F$ of $S$ such that $S \cap F = C$ and every refinement of $S$ consists of a refinement of $F$ coned to the points of $S \setminus C$. Moreover, all cells of $T$ containing $C$ must have $F$ refined the same way, so that there is a bijection between the refinements of $T$ and the polyhedral subdivisions of $C$, as a point set. The result then follows from the fact (see below) that a minimal affinely dependent set $C$ has exactly three subdivisions: the trivial one and two triangulations. □

[7] Although with one notable difference. For a general vector configuration not every function $w\colon \mathcal{A} \to \mathbb{R}$ produces a lift with a well-defined "lower envelope". Only the functions that do, namely those for which a linear hyperplane exists containing or lying below all the lifted vectors, define a regular polyhedral subdivision. These functions form a cone in $L_{\mathcal{A}}$. The secondary fan is still well-defined but, of course, it cannot be the normal fan of a polytope. It is, however, the normal fan of an unbounded convex polyhedron, called the *secondary polyhedron* of $\mathcal{A}$ [12].

This lemma allows us to understand a flip, even in the non-regular case, as a relation or a transformation between its two refinements. This is the usual usage of the word "flip", and our next topic.

**Flips as elementary changes.** A minimal affinely dependent set $C$ is called a *circuit* in geometric combinatorics. The points in a circuit $C = \{c_1, \ldots, c_k\}$ satisfy a unique (up to a constant) affine dependence equation $\lambda_1 c_1 + \cdots + \lambda_k c_k = 0$ with $\sum \lambda_i = 0$, and all the $\lambda_i$ must be non zero (or otherwise $C$ is not minimally dependent). This affine dependence implicitly decomposes $C$ into two subsets

$$C_+ = \{c_i : \lambda_i > 0\}, \quad C_- = \{c_i : \lambda_i < 0\}.$$

The pair $(C_+, C_-)$ is usually called a *signed* or *oriented* circuit. We will slightly abuse notation and speak of "the circuit $C = (C_+, C_-)$", unless we need to emphasize the distinction between the set $C$ (the *support* of the circuit) and its partition.

A more geometric description is that $(C_+, C_-)$ is the only partition of $C$ into two subsets whose convex hulls intersect, and that they intersect in their relative interiors. This is usually called *Radon's property* [58] and the oriented circuit a *Radon partition*.

Spanning and affinely independent subsets of $C$ are all the sets of the form $C \setminus \{c_i\}$. Moreover, by Radon's property two such sets $C \setminus \{c_i\}$ and $C \setminus \{c_j\}$ can be cells in the same triangulation of $C$ if and only if $c_i$ and $c_j$ lie in the same side of the partition. In other words:

**Lemma 1.13.** *A circuit $C = (C_+, C_-)$ has exactly two triangulations:*

$$T_+^C := \{C \setminus \{c_i\} : c_i \in C_+\}, \quad T_-^C := \{C \setminus \{c_i\} : c_i \in C_-\}.$$

This leads to a second definition of flip, equivalent to Definition 1.11, but more operational. This is the definition originally devised by Gel'fand, Kapranov and Zelevinskii [29]. The *link* of a set $\tau \subseteq \mathcal{A}$ in a triangulation $T$ of $\mathcal{A}$ is defined as

$$\operatorname{link}_T(\tau) := \{\rho \subseteq \mathcal{A} : \rho \cap \tau = \emptyset, \rho \cup \tau \in T\}.$$

**Definition 1.14.** Let $T_1$ be a triangulation of a point set $\mathcal{A}$. Suppose that $T_1$ contains one of the triangulations, say $T_+^C$, of a circuit $C = (C_+, C_-)$. Suppose also that all cells $\tau \in T_+^C$ have the same link in $T_1$, and call it $L$.

Then, we say that $C$ *supports a geometric bistellar flip* (or a *flip*, for short) in $T_1$ and that the following triangulation $T_2$ of $\mathcal{A}$ is obtained from $T_1$ by this flip:

$$T_2 := T_1 \setminus \{\rho \cap \tau : \rho \in L, \tau \in T_+^C\} \cup \{\rho \cap \tau : \rho \in L, \tau \in T_-^C\}.$$

If $i = |C_+|$ and $j = |C_-|$ we say that the flip is of type $(i, j)$. Flips of types $(1, j)$ and $(i, 1)$ are called, *insertion* and *deletion* flips, since they add or remove a vertex in the triangulation.

The *graph of flips* of $\mathcal{A}$ has as vertices all the triangulations of $\mathcal{A}$ and as edges the geometric bistellar flips between them.

Of course, an $(i, j)$ flip can always be reversed, giving a $(j, i)$ flip. The reason for the words "geometric bistellar" in our flips can be found in Section 2.2.

**Example 1.15** (Examples 1.2 and 1.7 continued). The change between the two top triangulations in Figure 3 is a $(2, 2)$ flip, as is the change between the two bottom ones. The flip from the top-right to the bottom-right is a $(1, 3)$ flip ("1 triangle disappears and 3 are inserted") and the flip from the top-left to the bottom-left is a $(1, 2)$ flip ("one edge is removed, together with its link, and two are inserted, with the same link"). The latter is supported in the circuit formed by the three collinear points.

We omit the proof of the following natural statement.

**Theorem 1.16.** *Definitions* 1.11 *and* 1.14 *are equivalent: two triangulations $T_1$ and $T_2$ of a point set $\mathcal{A}$ are connected by a flip in the sense of* 1.14 *if and only if they are the two proper refinements of a flip in the sense of* 1.11.

The following two facts are proved in [65]:

**Remark 1.17.** 1. If all proper refinements of a subdivision $T$ are triangulations, then $T$ has exactly two of them and $T$ is a flip. That is to say, flips are exactly the "next-to-minimal" elements in the refinement poset of all subdivisions of $\mathcal{A}$.

2. Every non-regular subdivision can be refined to a non-regular triangulation. In particular, not only edges of the secondary polytope correspond to flips between two regular triangulations, but also every flip between two regular triangulations corresponds to an edge.

**Detecting flips.** Definitions 1.11 and 1.14 are both based on the existence of a *flippable circuit* $C$ with certain properties. But in order to detect flips only some circuits need to be checked:

**Lemma 1.18.** *Every flip in a triangulation $T$ other than an insertion flip is supported in a circuit contained in the union of two adjacent cells of $T$.*

Observe that the circuit contained in two adjacent cells always exists and is unique. Also, that the insertion flips left aside in this statement are easy to detect:[8] There is one for each point $a \in \mathcal{A}$ not used in $T$, that inserts the point $a$ by subdividing the minimum (perhaps not full-dimensional) simplex $\tau \subseteq \sigma \in T$ such that $a \in \text{conv}(\tau)$. The flippable circuit is $(\{a\}, \tau)$.

*Proof.* Let $C = (C_+, C_-)$ be a circuit that supports a flip in $T$, with $|C_+| \geq 2$. Observe that $|C_+|$ is also the number of many maximal simplices in $T_+^C$, so let $\tau_1$ and $\tau_2$ be two of them, which differ in a single element, and let $\rho$ be an element of $\text{link}_T(\tau_1) = \text{link}_T(\tau_2)$. Then, $\rho \cup \tau_1$ and $\rho \cup \tau_2$ are adjacent cells in $T$ and $C$ is the unique circuit contained in $\tau_1 \cup \tau_2 \cup \rho$. □

[8] We mean, theoretically. Algorithmically, insertion flips are far from trivial since they imply locating the simplex of $T$ that contains the point $a$ to be inserted, which takes about the logarithm of the number of simplices in $T$. This is very expensive, since algorithms in computational geometry that use flipping in triangulations usually are designed to take constant time per flip other than an insertion flip. See Section 2.1.

**Monotone sequences of flips.** The graph of flips among regular triangulations of a point set $\mathcal{A}$ of dimension $d$ is connected, since it is the graph of a polytope.[9] A fundamental fact exploited in computational geometry is that one can actually flip between regular triangulations *monotonically*, in the following sense.

Let $w\colon \mathcal{A} \to \mathbb{R}$ be a certain generic lifting function. We can use $w$ to lift every triangulation $T$ of $\mathcal{A}$ as a function $H_{T,w}\colon \operatorname{conv}(\mathcal{A}) \to \mathbb{R}$, by affinely interpolating $w$ in each cell of $T$. We say that $T_1 <_w T_2$ ("$T_1$ is below $T_2$, with respect to $w$") if $H_{T_1,w} \leq H_{T_2,w}$ pointwise and $H_{T_1,w} \neq H_{T_2,w}$ globally. This defines a partial order $<_w$ on the set of all triangulations, whose global minimum and maximum are, respectively, $T_w$ and $T_{-w}$.[10]

**Definition 1.19.** A sequence of flips is monotone with respect to $w$ if every flip goes from a triangulation $T$ to a triangulation $T' <_w T$.

By definition of the secondary polytope $\Sigma(\mathcal{A})$ as having the secondary fan as its normal fan, lifting functions are linear functionals on it. Then, it is no surprise that for the regular triangulations $T_1$ and $T_2$ corresponding to vertices $v_{T_1}$ and $v_{T_2}$ of the secondary polytope one has[11]:

$$T_1 <_w T_2 \;\Rightarrow\; \langle w, v_{T_1}\rangle < \langle w, v_{T_1}\rangle.$$

In fact, $\langle w, v_T\rangle$ equals the volume between the graphs of the functions $H_{T_w,w}$ and $H_{T,w}$. Since the converse implication holds whenever $T_1$ and $T_2$ are related by a flip, we have:

**Lemma 1.20.** *For every lifting function $w$ and every regular triangulation $T$ there is a $w$-monotone sequence of flips from $T$ to the regular triangulation $T_w$.*

If $T$ is not regular this may be false, even in dimension 2:

**Example 1.21** (Examples 1.4 and 1.8 continued). Let $\mathcal{A}$ be the point configuration of Example 1.4 (see Figure 2), except perturbed by slightly rotating the interior triangle "123" counter-clockwise. That is,

$$\mathcal{A} = \begin{array}{c} \begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{matrix} \\ \begin{pmatrix} 4-\varepsilon & 0 & \varepsilon & 2 & 1 & 1 \\ \varepsilon & 4-\varepsilon & 0 & 1 & 2 & 1 \\ 0 & \varepsilon & 4-\varepsilon & 1 & 1 & 2 \end{pmatrix} \end{array},$$

[9] Even more, it is $(|\mathcal{A}| - d - 1)$-connected. Remember that a graph is called $k$-connected if removing less than $k$ vertices from it it stays connected. A classical theorem of Balinski [79] says that the graph of a $k$-polytope is $k$-connected.

[10] In case they are triangulations. If not, every triangulation that refines $T_w$ or $T_{-w}$ is, respectively, minimal or maximal.

[11] The same is true for non-regular triangulations. The point $v_T$ is well-defined, via the Gel'fand-Kapranov-Zelevinskii coordinates for the secondary polytope, even if $T$ is not regular. The only difference is that if $T$ is not regular then $v_T$ is not a vertex of the secondary polytope.

for a small $\varepsilon > 0$. This perturbation keeps the triangulation $T_1$ non-regular and makes $T_2$ regular. Let $w\colon \mathcal{A} \to \mathbb{R}$ lift the exterior triangle 123 to height zero and the interior triangle 456 to height one. The graph of $H_{T_2,w}$ is a strictly concave surface (that is, $T_2 = T_{-w}$) and there is no $w$-monotone flip in $T_1$, since its only three flips are the diagonal-edge flips on "16", "24" and "35", which are "towards $H_{T_2,w}$". This example appeared in [26].

Another explanation of why no $w$-monotone flip exists in $T_1$ is that when we close the graph of the function $H_{T_1,w}$ by adding to it the triangle 123, it becomes a non-convex polyhedron $P$ with the property that no tetrahedron (with vertices contained in those of $P$) is completely contained in the region enclosed by $P$. This polyhedron is affinely equivalent to Schönhardt's [69] classical example of a non-convex polyhedron in $\mathbb{R}^3$ that cannot be triangulated without additional vertices.

**1.3. The cases of small dimension or few points.** Throughout this section $\mathcal{A}$ denotes a point set with $n$ elements and dimension $d$.

**Sets with few points.** If $n = d+1$, then $\mathcal{A}$ is independent and the trivial subdivision is its unique triangulation. If $n = d + 2$ then $\mathcal{A}$ has a unique circuit and exactly two triangulations, connected by a flip. If $n = d + 3$, it was proved by Lee [44] that all triangulations are regular. Since the secondary fan is 2-dimensional, the secondary polytope is a polygon, whose graph (a cycle) is the graph of flips. If $n = d+4$, then $\mathcal{A}$ can have non-regular triangulations (see Example 1.4). Still, it is proven in [7] that every triangulation has at least three flips and that the flip-graph is 3-connected.

For point sets with $n = d + 5$ the flip-graph is not known to be always connected.

**Dimension 1.** Triangulating a one-dimensional point set is just choosing which of the interior points are used. That is, $n$ points in dimension 1 have $2^{n-2}$ triangulations. The flip-graph is the graph of an $(n-2)$-dimensional cube and all triangulations are regular. The secondary polytope is the same cube.

**Dimension 2.** In dimension two the graph of $(2, 2)$-*flips among triangulations using all points of* $\mathcal{A}$ [12] is known to be connected since long [40], and connectivity of the whole graph—including the triangulations that do not use all points and the insertion or deletion flips—is straightforward from that. Even more, one can always flip monotonically[13] from any triangulation to the Delaunay triangulation using only $(2, 2)$ flips. Quadratically many (with respect to the number of points) flips are sometimes necessary and always suffice (see, e.g., [25, p. 11]).

However, with general flips:

**Proposition 1.22.** *The flip-graph of any* $\mathcal{A} \subseteq \mathbb{R}^2$ *has diameter smaller than* $4n$.

[12] This is the graph usually considered in two-dimensional computational geometry literature.

[13] With respect to the lift $w(a) := \|a\|^2$.

*Proof.* Let $a$ be an extremal point of $\mathcal{A}$ and $T$ an arbitrary triangulation. If $T$ has triangles not incident to $a$ then there is at least a flip that decreases the number of them (proof left to the reader). Since the number of triangles in a planar triangulation with $v_i$ interior vertices and $v_b$ boundary vertices is exactly $2v_i + v_b - 2$ (by Euler's formula) we can flip from any triangulation to one with every triangle incident to $a$ in at most $2v_i + v_b - 3 < 2n - n_b$ flips.

Now, exactly as in the 1-dimensional case, the graph of flips between triangulations in which every triangle is incident to $a$ is the graph of a cube of dimension equal to the number of "boundary but non-extremal" points of $\mathcal{A}$. Hence, we can flip between any two triangulations in $(2n - n_b) + n_b + (2n - n_b) < 4n$ flips. □

**Remark 1.23.** The preceding proof is another example of monotone flipping, this time with respect to any lifting function $w\colon \mathcal{A} \to \mathbb{R}$ with $w(a) << w(b)$, for all $b \in \mathcal{A} \setminus \{a\}$. In essence, this lifting produces the so-called *pulling* triangulation of $\mathcal{A}$. More precisely, for a point set $\mathcal{A}$ in arbitrary dimension and a given ordering $a_1, \ldots, a_n$ of the points in $\mathcal{A}$ one defines [17], [44], [45], [79]:

- The *pulling triangulation* of $\mathcal{A}$, as the regular triangulation given by the lift $w(a_i) := -t^i$, for a sufficiently big constant $t \in \mathbb{R}$. It can be recursively constructed as the triangulation that joins the last point $a_n$ to the pulling triangulation of every facet of $\text{conv}(\mathcal{A})$ that does not contain $a_n$.
- The *pushing triangulation* of $\mathcal{A}$, as the regular triangulation given by the lift $w(a_i) := t^i$, for a sufficiently big constant $t \in \mathbb{R}$. It can be recursively constructed as the triangulation that contains $T_{n-1}$ and joins $a_n$ to the part of the boundary of $T_{n-1}$ visible from $a_n$, where $T_{n-1}$ is the pushing triangulation of $\mathcal{A} \setminus \{a_n\}$.

Pushing and pulling triangulations are examples of *lexicographic* triangulations, defined by the lifts $w(a_i) := \pm t^i$ for sufficiently big $t$.

Summing up, monotone flipping in the plane (a) works even for non-regular triangulations if the "objective function" $w$ is either the Delaunay or a lexicographic one (the proof for the pushing case is left to the reader); (b) gives a linear sequence of flips for the pulling case, but may produce a quadratic one for the Delaunay case; (c) does not work for arbitrary $w$ (Example 1.21).

Let us also mention that in dimension two every triangulation is known to have at least $n-3$ flips [23] (the dimension of the secondary polytope), and at least $\lceil (n-4)/2 \rceil$ of them of type (2, 2) [35]. The flip-graph is not known to be $(n-3)$-connected.

**Dimension 3.** Things start to get complicated:

If $\mathcal{A}$ is in convex position [14] then every triangulation of it has at least $n-4$ flips [23], but otherwise $\mathcal{A}$ can have flip-deficient triangulations.[15] The smallest possible example, with eight points, is described in [7], based on Example 1.4. Actually, for every $n$

[14] *Convex position* means that all points are vertices of $\text{conv}(\mathcal{A})$.

there are triangulations with essentially $n^2$ vertices and only $O(n)$ flips [63]. This is true even in general position.[16]

The flip-graph is not known to be connected, even if $\mathcal{A}$ is in convex and general position. The main obstacle to proving connectivity (in case it holds!) is probably that one cannot, in general, *monotonically* flip to either the Delaunay, the pushing, or the pulling triangulations. For the Delaunay triangulation this was shown in [37]. For the other two we describe here an example.

**Example 1.24** (Examples 1.4, 1.8 and 1.21 continued). Let $\mathcal{A}$ consist of the following eight points in dimension three:

$$\mathcal{A} = \begin{array}{c} \begin{array}{cccccccc} a_1 & a_2 & a_3 & b_1 & b_2 & b_3 & c_1 & c_2 \end{array} \\ \begin{pmatrix} 4-\varepsilon & 0 & \varepsilon & 2 & 1 & 1 & 4/3 & 4/3 \\ \varepsilon & 4-\varepsilon & 0 & 1 & 2 & 1 & 4/3 & 4/3 \\ 0 & \varepsilon & 4-\varepsilon & 1 & 1 & 2 & 4/3 & 4/3 \\ 0 & 0 & 0 & 1 & 1 & 1 & 10 & -10 \end{pmatrix} \end{array},$$

The first six points are exactly the (lifted) point set of Example 1.21, and have the property that no tetrahedron with vertices contained in these six points is contained in the non-convex Schönhardt polyhedron $P$ having as boundary triangles $\{a_1, a_2, a_3\}$, $\{b_1, b_2, b_3\}$, $\{a_i, a_{i+1}, b_i\}$ and $\{a_{i+1}, b_i, b_{i+1}\}$ (the latter for the three values of $i$, and with indices regarded modulo three). The last two points $c_1$ and $c_2$ of the configuration lie far above and far below this polyhedron. $c_1$ sees every face of $P$ except the big triangle $\{a_1, a_2, a_3\}$, while $c_2$ sees only this triangle.

Let $T$ be the triangulation $T$ of $\mathcal{A}$ obtained removing the big triangle from the boundary of $P$, and joining the other seven triangles to both $c_1$ and $c_2$. We leave it to the reader to check that there is no monotone sequence of flips towards the pushing triangulation with respect to any ordering ending in $c_2$, and there is no monotone sequence of flips towards the pulling triangulation with respect to any ordering ending in $c_1$.

**Higher dimension.** There are the following known examples of "bad behavior":

- In *dimension four*, there are triangulations with arbitrarily many vertices and a bounded number of flips [63]. They are constructed adding several layers of "the same" triangulated 3-sphere one after another.
- In *dimension five*, there are point sets with a disconnected graph of triangulations [67]. The smallest one known has 26 points, but one with 50 points is easier to describe: It is the Cartesian product of $\{0, 1\}$ with the vertex set and the centroid of a regular 24-cell.

[15] We say a triangulation is *flip-deficient* if it has less than $n - d - 1$ flips; that is, less than the dimension of the secondary polytope.

[16] Although this is not mentioned in [63], the construction there can be perturbed without a significant addition of flips.

- In *dimension six*, there are triangulations without flips at all [64]. The example is again a cartesian product, now of a very simple configuration of four points in $\mathbb{R}^2$ and a not-so-simple (although related to the 24-cell too) configuration of 81 points in $\mathbb{R}^4$. There are also point sets *in general position* with a disconnected graph of triangulations (Section 3 of this paper and [68]). Only 17 points are needed.

## 2. The context

**2.1. Bistellar flips and computational geometry.** The first and most frequently considered flips in computational geometry are $(2, 2)$ flips in 2-dimensional point sets. Seminal papers of Lawson [40], [41] prove that every triangulation can be monotonically transformed to the Delaunay triangulation by a sequence of $O(n^2)$ such flips.

Lawson himself, in 1986 [42], is close to defining flips in arbitrary dimension, even in the case of special position. Around 1990,[17] B. Joe realizes that in dimension three one cannot, in general, monotonically flip from any triangulation to the Delaunay triangulation [37] but, still, the following incremental algorithm works [38]: insert the points one by one, each by an insertion flip in the Delaunay triangulation of the already inserted points. After each insertion, monotonically flip to the new Delaunay triangulation by flips that increase the star of the inserted point.

V. T. Rajan [59] does essentially the same in arbitrary dimension and Edelsbrunner and Shah [26], already aware at least partially of the theory of secondary polytopes, generalize this to flipping towards the regular triangulation $T_w$ by $w$-monotone flips, for an arbitrary $w$.

If one disregards the efficiency of the algorithm, the main result of [26] follows easily from Lemma 1.20. But efficiency is the main point in computational geometry, and one of the important features in [38] and [26] is to show that the sequence of flips can be found and performed spending constant time per flip (in fixed dimension). An exception to this time bound are the insertion steps. Theoretically, they are just another case of flip. But in the algorithm they have a totally different role since they involve locating where the new point needs to be inserted. To get good time bounds for the location step, the standard incremental algorithm is "randomized",[18] and it is proved that the total *expected* time taken by the $n$ insertion steps is bounded above by $O(n \log n)$ in the plane and $O(n^{\lceil d/2 \rceil})$ in higher dimension. The latter is the same as the worst-case size of the Delaunay triangulation, or actually of any triangulation.

This incremental-randomized-flipping method can be considered the standard algorithm for the Delaunay triangulation in current computational geometry. It is the only one described in the textbooks [21] and [25]. In the survey [8], it is the first of

[17] Birth year of secondary polytopes and geometric bistellar flips as we have defined them [28].

[18] That is, the ordering in which the points are inserted is considered random among the $n!$ possible orderings. This trick was first introduced in [19] for convex hulls, then applied to 2-D Delaunay triangulations in [32].

four described in the plane but the only one detailed in dimension three, as "the most intuitive and easy to implement".

**Remark 2.1.** Computational geometry literature normally only considers full-dimensional flips; that is, flips of type $(i, j)$ with $i + j = d + 2$. In particular, [8], [21], [25], [26] and [38] describe the incremental flipping algorithm only for point sets in general position. The only mention in those references to the effect of allowing special position in the flipping process seems to be that, according to [25], for the six vertices of a regular octahedron *"none of the three tetrahedrizations permits the application of a two-to-three or a three-to-two flip. The flip graph thus consists of three isolated nodes"*.

However, with the general definition of flip the incremental-flipping algorithm can be directly applied to point sets in special position, as done recently by Shewchuk [70]. Shewchuk's algorithm actually computes the so-called *constrained regular triangulation* of the point set for any lift $w$ and constraining complex $K$. This is defined as the unique[19] triangulation $T$ containing $K$ and in which every simplex of $T \setminus K$ is lifted by $w$ to have a locally convex star.[20]

**2.2. Bistellar flips and combinatorial topology.** Bistellar flips can be defined at a purely combinatorial level, for an abstract simplicial complex. Let $\Delta$ be a simplicial complex, and let $\sigma \in \Delta$ be a simplex, of any dimension. The *stellar subdivision* on the simplex $\sigma$ is the simplicial complex obtained inserting a point in the relative interior of $\sigma$. This subdivides $\sigma$, and every simplex $\tau$ containing it, into $\dim \sigma + 1$ simplices of the same dimension. Two simplicial complexes $\Delta_1$ and $\Delta_2$ are said to differ in a *bistellar flip* if there are simplices $\sigma_1 \in \Delta_1$ and $\sigma_2 \in \Delta_2$ such that the stellar subdivisions of $\Delta_1$ and $\Delta_2$ on them produce the same simplicial complex. The bistellar operation from $\Delta_1$ to $\Delta_2$ is said to be of type $(i, j)$ if $i = \dim \sigma_1 + 1$ and $j = \dim \sigma_2 + 1$. Observe that geometric bistellar flips, as defined in Definition 1.14, are combinatorially bistellar flips.

Combinatorial bistellar flips have been proposed as an algorithmic tool for exploring the space of triangulations of a manifold [21] or to recognize the topological type of a simplicial manifold [18], [46]. In particular, Pachner [56] has shown that any two triangulations of PL-homeomorphic manifolds are connected by a sequence of topological bistellar flips. But for this connectivity result additional vertices are allowed to be inserted into the complex, via flips of type $(i, 1)$.

[19] If it exists, which is not always the case.

[20] Shewchuk's algorithm is incremental, treating the simplices in $K$ similarly to the points in the standard incremental algorithm: they are inserted one by one (in increasing order of dimension) and after each insertion the regular, constrained to the already added simplices, triangulation is updated using geometric bistellar flips. The algorithm's running time is $O(n^{\lfloor d/2 \rfloor + 1} \log n)$. The extra $\log n$ factor comes from a priority queue that is needed to decide in which order the flips are performed, to make sure that no "local optima" instead of the true constrained Delaunay triangulation, is reached. The extra $n$ factor (only in even dimension) is what randomization saves in the standard incremental-flipping algorithm. Randomization would not do the same here (Shewchuk, personal communication).

[21] Besides its intrinsic interest, this problem arises in quantum gravity modelization [3], [54].

The situation is much different if we do not allow insertion flips: Dougherty et al. [24] show that there is a topological triangulation of the 3-sphere, with 15 vertices, that does not admit any flip other than insertion flips.[22] If this triangulation was realizable geometrically in $\mathbb{R}^3$ (removing from the 3-sphere the interior of any particular tetrahedron) it would provide a triangulation in dimension three without any geometric bistellar flips. Unfortunately, Dougherty et al. show that it cannot be geometrically embedded.

**2.3. Bistellar flips and topological combinatorics.** A standard construction in topological combinatorics [16] is to associate to a poset $P$ its *order complex*: an abstract simplicial complex whose vertices are the elements of $P$ and whose simplices are the finite chains (totally ordered subsets) of $P$. In this sense one can speak of the topology of the poset. If the poset has a unique maximum (as is the case with the refinement poset of subdivisions of a point set $\mathcal{A}$) or minimum, one usually removes them or otherwise the order complex is trivially contractible (that is, homotopy equivalent to a point). This is what we mean when we say that the refinement poset of subdivisions of the point set of Section 3 is not connected.

The refinement poset of polyhedral subdivisions of $\mathcal{A}$ is usually called the *Baues poset* of $\mathcal{A}$ and its study is *the generalized Baues problem*. To be precise, Baues posets were introduced implicitly in [14] and explicitly in [13] in a more general situation where one has an affine projection $\pi$ from the vertex set of a polytope $P \in \mathbb{R}^{d'}$ to a lower dimensional affine space $\mathbb{R}^d$. In this general setting, one considers the point set $\mathcal{A} := \pi(\text{vertices}(P))$ and is interested in the polyhedral subdivisions of $\mathcal{A}$ that are compatible with $\pi$ in a certain sense (basically, that the preimage of every cell is the set of vertices of a face in $P$). In the special case were $P$ is a simplex (and hence $d' = n - 1$, where $n$ is the number of points in $\mathcal{A}$) every polyhedral subdivision is compatible. This is the case of primal interest in this paper, but there are at least the following two other cases that have attracted attention. (See [61] for a very complete account of different contexts in which Baues posets appear, and [79, Chapter 9] for a different treatment of the topic):

- When $P$ is a cube, its projection is a *zonotope* $Z$ and the $\pi$-compatible subdivisions are the *zonotopal tilings* of $Z$ [79]. The finest ones are *cubical tilings*, related by *cubical flips*.
- When $d = 1$ and $P$ is arbitrary, the $\pi$-compatible subdivisions are called *cellular strings*, since they correspond to monotone sequences of faces of $P$. The finest ones are *monotone paths* of edges and are related by *polygon flips*.

[22] Dougherty et al. only say that their triangulation does not have any $(3, 2)$, $(2, 3)$ or $(1, 4)$ flips, which are the "full-dimensional" types of flips. But their arguments prove that even considering degenerate flips, the only possible ones in their triangulation are insertion flips of type $(i, 1)$. Indeed, the two basic properties that their triangulations has are that (a) its graph is complete, which prevents flips of type $(3, 2)$, but also $(2, 2)$ and $(1, 2)$ and (b) no edge is incident to exactly three tetrahedra, which prevents flips of type $(1, 4)$ and $(2, 3)$, but also $(1, 3)$.

The name *Baues* for these posets comes from the fact that H. J. Baues was interested in their homotopy type in a very particular case (in which, among other things, $d = 1$) and conjectured it to be that of a sphere of dimension $d' - 2$ [9]. Billera et al. [13] proved this conjecture for all Baues posets with $d = 1$, and the conjecture that the same happened for arbitrary $d$ (with the dimension of the sphere being now $d' - d - 1$) became known as the *generalized Baues conjecture*. It was inspired by the fact that the *fiber polytope* associated to the projection $\pi$—a generalization of the secondary polytope, introduced in [14]—has dimension $d' - d$ and its face lattice is naturally embedded in the Baues poset.

Even after the conjecture in its full generality was disproved by a relatively simple example with $d' = 5$ and $d = 2$ [60], the cases where $P$ is either a simplex or a cube remained of interest. As we have said, the latter is disproved in the present paper for the first time. The former remains open and has connections to oriented matroid theory, as we now show.

Recall that the oriented matroid (or order type) of a point set $\mathcal{A}$ of dimension $d$ (or of a vector configuration of rank $d + 1$) is just the information contained in the map $\binom{\mathcal{A}}{d+1} \to \{-1, 0, +1\}$ that associates to each $(d + 1)$-element subset of $\mathcal{A}$ the sign of its determinant (that is, its orientation). But oriented matroids (see [17] as a general reference) are axiomatically defined structures which may or may not be realizable as the oriented matroids of a real configuration, in much the same way as, for example, a topological space may or may not be metrizable.

It turns out that the theory of triangulations of point and vector configurations generalizes nicely to the context of perhaps-non-realizable oriented matroids, with the role of regular triangulations being played by the so-called *lifting triangulations*: triangulations that can be defined by an oriented matroid lift (see [17, Section 9.6] or [66, Section 4]).

One of the basic facts in oriented matroid duality is that the lifts of an oriented matroid $\mathcal{M}$ are in bijection to the one-point extensions of its dual $\mathcal{M}^*$. In particular, the space of lifts of $\mathcal{M}$ equals the so-called extension space of the dual oriented matroid $\mathcal{M}^*$. Here, both the space of lifts and the space of extensions are defined as the simplicial complexes associated to the natural poset structures in the set of all lifts/extensions of the oriented matroid. This makes the following conjecture of Sturmfels and Ziegler [75] be relevant to this paper:

**Conjecture 2.2.** The extension space of a realizable oriented matroid of rank $r$ is homotopy equivalent to a sphere of dimension $r - 1$.

The reader may be surprised that we call this a conjecture: if the extension space of an oriented matroid is the analogue of a secondary fan, should not the extension space of a realizable oriented matroid be automatically "a fan", hence a sphere? Well, no: even if an oriented matroid $\mathcal{M}$ is realizable, some of its extensions may not be realizable. Those will appear in the extension space. Even worse, if $\mathcal{M}$ is realized as a vector configuration $\mathcal{A}$, some realizable extensions of $\mathcal{M}$ may only be realizable as extensions of other realizations of $\mathcal{M}$. Actually, Sturmfels and Ziegler show that

the space of realizable extensions of a realizable oriented matroid *does not* in general have the homotopy type of a sphere!

**Example 2.3** (Example 1.4 continued). Consider the point configuration of Example 1.4 (two parallel triangles one inside the other). An additional point added to this configuration represents an extension of the underlying oriented matroid. In particular, there is an extension by a point collinear with each of the three pairs of corresponding vertices of the two triangles.

But any small perturbation of the point set gives another realization of the same oriented matroid, since the original point set is in general position. However, this perturbation will, in general, not keep the lines through those three pairs of vertices colliding. So, the extension we have described is no longer realizable as a geometric extension of the new realization.

There is a class of configurations specially interesting in this context: the so-called Lawrence polytopes. A *Lawrence oriented matroid* is an oriented matroid whose dual is centrally symmetric. Similarly, a *Lawrence polytope* is a polytope whose vertex set has a centrally symmetric *Gale transform*. There is essentially one Lawrence polytope associated to each and every realizable oriented matroid. The following result is a combination of a theorem of Bohne and Dress (see [79], for example) and one of the author of this paper [34], [66]:

**Theorem 2.4.** *Let $\mathcal{M}$ be a realizable oriented matroid and let $P$ be the associated Lawrence polytope. Then, the following three posets are isomorphic:*

1. *The refinement poset of polyhedral subdivisions of $P$.*
2. *The extension space of the* (*also realizable*) *dual oriented matroid $\mathcal{M}^*$.*
3. *The refinement poset of zonotopal tilings of the zonotope associated to* (*any realization of*) *$\mathcal{M}$.*

**Corollary 2.5.** *The following three statements are equivalent:*

1. *The generalized Baues conjecture for the polyhedral subdivisions of Lawrence polytopes.*
2. *The extension space conjecture for realizable oriented matroids.*
3. *The generalized Baues conjecture for the zonotopal tilings of zonotopes.*

Moreover, if $\mathcal{A}$ is a point configuration and $P$ its associated Lawrence polytope, then there is a surjective map between the poset of subdivisions of $P$ and the poset of *lifting* (in the oriented matroid sense) subdivisions of $\mathcal{A}$. This follows from that facts that "Lawrence polytopes only have lifting subdivisions" and "lifting subdivisions can be lifted to the Lawrence polytope", both proved in [66].

In particular, if the flip-graph of a certain point set $\mathcal{A}$ is not connected and has lifting triangulations in several connected components, then the graph of cubical flips

between zonotopal tilings of a certain zonotope is not connected either, thus answering question 1.3 in [61]. If, moreover, $\mathcal{A}$ is in general position, it would disprove the three statements in Corollary 2.5. We do not know whether the disconnected flip-graph in Section 3 has this property. The examples in [64], [67] are easily seen to be based in non-lifting triangulations.

**Remark 2.6.** The extension space conjecture is the case $k = d - 1$ of the following far-reaching conjecture by MacPherson, Mnëv and Ziegler [61, Conjecture 11]: that the poset of all strong images of rank $k$ of any realizable oriented matroid $\mathcal{M}$ of rank $d$ (the so-called *OM-Grassmannian of rank k* of $\mathcal{M}$) is homotopy equivalent to the real Grassmannian $G^k(\mathbb{R}^d)$. This conjecture is relevant in matroid bundle theory [5] and the *combinatorial differential geometry* of MacPherson [48].

An important achievement in this context is the recent result of Biss [15] proving this conjecture whenever $\mathcal{M}$ is a "free oriented matroid". In this case the OM-Grassmannian is the space of all oriented matroids of a given cardinality and rank, usually called the MacPhersonian. The result of Biss includes the case $n = \infty$ (in which the MacPhersonian is defined as a direct limit of all the MacPhersonians of a given rank) and implies that the theory of "oriented matroid bundles for combinatorial differential manifolds" developed by MacPherson [48] is equivalent to the theory of real vector bundles on real differential manifolds. A first, seminal, result in this direction was the "combinatorial formula" by Gel'fand and MacPherson for the Pontrjagin class of a triangulated manifold [30].

**2.4. Bistellar flips and algebraic geometry.** Bistellar flips are related to algebraic geometry from their very birth. Indeed, Definition 1.14, as well as that of secondary polytope and Theorem 1.6 were first given by Gel'fand, Kapranov and Zelevinskii during their study of discriminants of a sparse polynomial [28]. By a sparse polynomial we mean, here, a multivariate polynomial $f$ whose coefficients are considered parameters but whose set of (exponent vectors of) monomials is a fixed point set $\mathcal{A} \subseteq \mathbb{Z}^d$. Gel'fand, Kapranov and Zelevinskii prove that the secondary polytope of $\mathcal{A}$ equals the Newton polytope of the Chow polynomial of $f$, where the Chow polynomial is a certain resultant defined in terms of $f$. Similarly, the secondary polytope is related to the discriminant of $f$ (the $\mathcal{A}$-discriminant) although a bit less directly: it is a Minkowski summand of the Newton polytope of the $\mathcal{A}$-discriminant.

A stronger, and more classical, relation between triangulations of point sets and algebraic geometry comes from the theory of toric varieties [27], [55]. As is well-known, every rational convex polyhedral fan $\Sigma$ (in our language, every polyhedral subdivision of a rational vector configuration) has an associated toric variety $X_\Sigma$, of the same dimension. $X_\Sigma$ is non-singular if and only if $\Sigma$ is simplicial (i.e., a triangulation) and unimodular. The latter means that every cone is spanned by integer vectors with determinant $\pm 1$. If $\Sigma$ is a non-unimodular triangulation, then $X_\Sigma$ is an orbifold; that is, it has only quotient singularities.

A stellar subdivision, that is, an insertion flip, in $\Sigma$ corresponds to an equivariant

blow-up in $X_\Sigma$. Hence, a deletion flip produces a blow-down and a general flip produces a blow-up followed by a blow down. In this sense, the connectivity question for triangulations of a vector configuration is closely related to the following result, conjectured by Oda [51] and proved by Morelli and Włodarczyk [52], [78].[23]

**Theorem 2.7.** *Every proper and equivariant birational map $f : X_\Sigma \to X_{\Sigma'}$ between two nonsingular toric varieties can be factorized into a sequence of blowups and blowdowns with centers being smooth closed orbits* (*weak Oda's conjecture*).

More precisely, Oda's conjecture, in its weak form, is equivalent to saying that every pair of unimodular simplicial fans can be connected by a sequence of bistellar flips passing only through unimodular fans (and, actually, it is proved this way). But observe that in this result the set of vectors allowed to be used is not fixed in advance: additional ones are allowed to be flipped-in and eventually flipped-out. Our construction in [67] actually shows that the result is not true if we do not allow for extra vectors to be inserted.

The relation of the graph of flips to toric geometry is even closer if one looks at certain schemes associated to a toric variety. In order to define them we first look at secondary polytopes in a different way, as a particular case of fiber polytopes [14]:

Assume that $\mathcal{A}$ is an integer point configuration and let $\Delta$ be the unit simplex of dimension $|\mathcal{A}| - 1$ in $\mathbb{R}^{|\mathcal{A}|}$. Let $Q = \text{conv}(\mathcal{A})$ and let $\pi : \Delta \to Q$ be the affine projection sending the vertices of $\Delta$ to $\mathcal{A}$. The chamber complex of $\mathcal{A}$ is the coarsest common refinement of all its triangulations. It is a polyhedral complex with the property that for any $b$ and $b'$ in the same *chamber* the fibers $\pi^{-1}(b)$ and $\pi^{-1}(b')$ are polytopes with the same normal fan.

**Theorem 2.8** (Billera et al. [14]). *The secondary polytope of $\mathcal{A}$ equals the Minkowski integral of $\pi^{-1}(b)$ over $Q$.*

Combinatorially, then, the secondary polytope of $\mathcal{A}$ equals the Minkowski sum of a finite number of $\pi^{-1}(b)$'s, with one $b$ chosen in each chamber.

Now, for each $b \in Q$, consider the toric variety associated to the normal fan of the fiber $\pi^{-1}(b)$. Since the normal fan is the same whenever $b$ and $b'$ lie in (the relative interior of) the same cell of the chamber complex, we denote this toric variety $V_\sigma$, where $\sigma$ is a cell (of any dimension) of the chamber complex. If $b \in \sigma$ and $b' \in \tau$ for two chambers with $\tau \subseteq \overline{\sigma}$ then the normal fan of $\pi^{-1}(b)$ refines the normal fan of $\pi^{-1}(b')$, which implies that there is a natural equivariant morphism $f_{\sigma\tau} : V_\sigma \to V_\tau$. We finally denote $\Lambda_{\mathcal{A}} := \varprojlim V_\sigma$ the inverse limit of all the $V_\sigma$ and morphisms $V_{\sigma\tau}$. It has the following two interpretations:

1. Let $X_\Delta$ be the projective space of dimension $|\mathcal{A}| - 1$, which is the toric variety associated with the simplex $\Delta$ (what follows is valid for any polytope $\Delta$). The

[23] Morelli's paper [52] claimed to have proved the following: that we can insist on the sequence to consist of first a sequence of only blowups and then one of only blowdowns (strong Oda's conjecture). Some errors were found in this part of his paper [2, 53] and, according to [1], the strong conjecture is still open, even in dimension three.

toric varieties $V_\sigma$ are the different toric geometric invariant theory quotients of $X_\Delta$ modulo the algebraic sub-torus whose characters are the monomials with exponents in $\mathcal{A}$ [39, Section 3]. $\Lambda_\mathcal{A}$ is the inverse limit of all of them, which contains the Chow quotient as an irreducible component [39, Section 4].

2. In [4], Alexeev is interested, among other things, in the moduli space $M$ of *stable semi-abelic toric pairs* for an integer polytope $Q$ (see Sections 1.1.A and 1.2.B in [4] for the definitions). The author shows that there is a finite morphism $M \to \Lambda_\mathcal{A}$ (Corollary 2.11.11), where $\mathcal{A}$ is the set of all integer points in $Q$, and uses $\Lambda_\mathcal{A}$ (that he denotes $M_{\text{simp}}$) as a simplified model for studying $M$.

   Although there $\mathcal{A}$ is assumed to be the set of all lattice points in a polytope, the connection of $\Lambda_\mathcal{A}$ with $\Sigma_c(\mathcal{A})$ carried out in the proof of the following theorem is independent of this fact.

**Theorem 2.9.** *The scheme $\Lambda_\mathcal{A}$ is connected if and only if the graph of triangulations of $\mathcal{A}$ is connected.*

*Proof* (Sketch). Alexeev introduces the following poset structure on the set of all polyhedral subdivisions of $\mathcal{A}$: Given two subdivisions $S_1$ and $S_2$ we consider $S_1 < S_2$ if: (a) $S_1$ refines $S_2$, (b) the restriction of $S_1$ to each cell $B$ of $S_2$ is a regular subdivision $S_B$ of $B$, and (c) the lifting functions of the regular subdivisions of cells of $S_2$ can be chosen so that the restrictions of them to common faces of cells differ by an affine function.

This poset is called the "coherent poset of subdivisions of $\mathcal{A}$" in [64], to distinguish it from the usual poset of subdivisions, where only the first condition (refinement) is imposed. Then, he shows that the scheme $\Lambda_\mathcal{A}$ is connected if and only if the coherent refinement poset is connected. (More precisely, he shows that there is a natural moment map defined on $\Lambda_\mathcal{A}$ whose image is the topological model of the poset). In turn, it is proven in [64] that the coherent refinement poset is connected if and only if the graph of triangulations of $\mathcal{A}$ is connected. □

A second scheme that relates triangulations and toric geometry is precisely the so-called toric Hilbert scheme. The toric ideal $I_\mathcal{A} \subseteq K[x_1, \ldots, x_n]$ associated to $\mathcal{A} = \{a_1, \ldots, a_n\} \in \mathbb{R}^d$ is generated by the binomials

$$\left\{x^\lambda - x^\mu : \lambda, \mu \in \mathbb{N}^n,\ \sum \lambda_i a_i = \sum \mu_i a_i\right\}.$$

Here, $x^\lambda := x_1^{\lambda_1} \cdots x_n^{\lambda_n}$. In other words, $I_\mathcal{A}$ is the lattice ideal of the lattice of integer affine dependences among $\mathcal{A}$. $\mathcal{A}$ defines the following $\mathcal{A}$-grading of monomials in $K[x_1, \ldots, x_n]$: the $A$-degree of $x^\lambda$ is the vector $x_1^{\lambda_1} \cdots x_n^{\lambda_n} \in \mathbb{Z}^d$. Of course, $I_\mathcal{A}$ is homogeneous with respect to this grading.

If $I$ is another $\mathcal{A}$-homogeneous ideal, the Hilbert function of $I$ is the map $\mathbb{Z}^d \to \mathbb{N}$ defined by $b \mapsto \dim_K I_b$ where $I_b$ is the part of $I$ of degree $b$. The *toric Hilbert scheme*

of $\mathcal{A}$ consists, as a set, of all the $\mathcal{A}$-homogenous ideals with the same Hilbert function as the toric ideal $I_{\mathcal{A}}$. It contains $I_{\mathcal{A}}$ as well as all its initial ideals, which form an irreducible component in its scheme structure.

The toric Hilbert scheme was introduced by Sturmfels in [73] (see also [74]) although its scheme structure was explicited later by Peeava and Stillman [57], who ask whether non-connected toric Hilbert schemes exist.

Sturmfels shows, among other things, that there is a natural map from the toric Hilbert scheme to the set of polyhedral subdivisions of $\mathcal{A}$. Moreover, the map is continuous when the latter is given either the poset topology or the "coherent poset topology" introduced in the proof of Theorem 2.9. The map is not surjective in general, so disconnected graphs of triangulations do not automatically imply disconnected Hilbert schemes.[24] However, Maclagan and Thomas [47], modifying the arguments of Theorem 2.9, show that the image of the map contains at least al the unimodular triangulations of $\mathcal{A}$. In particular:

**Corollary 2.10.** *If the graph of triangulations of an integer point configuration $\mathcal{A}$ is not connected and contains unimodular triangulations in non-regular connected components, then the toric Hilbert scheme of $\mathcal{A}$ is not connected.*

The example in [67] satisfies the hypothesis of this corollary. Hence:

**Theorem 2.11** (Santos [67]). *Let $\mathcal{A}_{50} \subset \mathbb{R}^5$ be the point set $\mathcal{A}_{25} \times \{0, 1\}$ where $\mathcal{A}_{25} \subset \mathbb{R}^4$ consists of the centroid and the 24 vertices of a regular 24-cell. The toric Hilbert scheme of $\mathcal{A}$ and the scheme $\Lambda_A$ defined above are both non-connected. They have at least 13 connected components, each with at least $3^{48}$ torus-fixed points.*

## 3. A construction

Let $\mathcal{A}(t) \subset \mathbb{R}^6$ be the point set defined by the columns of the following matrix, where $t$ is a positive real number. The matrix is written in two pieces for typographic reasons. As usual, the first row is just a homogenization coordinate:

$$\mathcal{A}(t) := \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{ccccccccc} O & a_1^+(t) & a_2^+(t) & a_3^+(t) & a_4^+(t) & a_5^+(t) & a_6^+(t) & a_7^+(t) & a_8^+(t) \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & -t & 0 & 0 & 1 & t & 0 & 0 \\ 0 & t & 1 & 0 & 0 & -t & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -t & 0 & 0 & 1 & t \\ 0 & 0 & 0 & t & 1 & 0 & 0 & -t & 1 \\ 0 & \sqrt{2} & 1 & 0 & -1 & -\sqrt{2} & -1 & 0 & 1 \\ 0 & 0 & 1 & \sqrt{2} & 1 & 0 & -1 & -\sqrt{2} & -1 \end{pmatrix} \end{array} \cdots$$

[24] Haiman and Sturmfels [33] have shown that this map factors as a morphism from the toric Hilbert scheme to the scheme $\Lambda_{\mathcal{A}}$ of the previous discussion, followed by the natural map from that scheme to the poset of subdivisions. The first map is the non-surjective one.

$$\cdots \begin{array}{c} \begin{matrix} a_1^-(t) & a_2^-(t) & a_3^-(t) & a_4^-(t) & a_5^-(t) & a_6^-(t) & a_7^-(t) & a_8^-(t) \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & t & 0 & 0 & -1 & -t & 0 & 0 \\ -t & -1 & 0 & 0 & t & -1 & 0 & 0 \\ 0 & 0 & -1 & t & 0 & 0 & -1 & -t \\ 0 & 0 & -t & -1 & 0 & 0 & t & -1 \\ \sqrt{2} & 1 & 0 & -1 & -\sqrt{2} & -1 & 0 & 1 \\ 0 & 1 & \sqrt{2} & 1 & 0 & -1 & -\sqrt{2} & -1 \end{pmatrix} \end{array}.$$

$\mathcal{A}(t)$ is not in general position. For example, for every $i = 1, 2, 3, 4$ we have:

$$a_i^+(t) + a_{i+4}^+(t) + a_i^-(t) + a_{i+4}^-(t) = 4O.$$

However, it is "sufficiently in general position" for the following to be true:

**Theorem 3.1.** *If $t$ is sufficiently small and $\mathcal{A}'(t)$ is any perturbation of $\mathcal{A}(t)$ in general position, then the graph of triangulations of $\mathcal{A}'(t)$ is not connected.*

When we say that a point set $\mathcal{A}'$ is a perturbation of another one $\mathcal{A}$ with the same cardinal $n$ and dimension $d$ we mean that all the determinants of $d+1$ points that are not zero in $\mathcal{A}$ keep their sign in $\mathcal{A}'$.[25] This concept also allows us to be precise as to how small do we need $t$ to be. Any $t$ such that $\mathcal{A}(t)$ is a perturbation of $\mathcal{A}(0)$ works.

The proof of Theorem 3.1 will appear in [68]. Here we only give a description of the combinatorics of $\mathcal{A}(t)$ and the ingredients that make the proof work. We look at $\mathcal{A}(0)$ first. In it:

- The projection to the first four coordinates $x_1, \ldots, x_4$ sends the eight pairs of points $\{a_i^+(t), a_{i+4}^+(t)\}$, and $\{a_i^-(t), a_{i+4}^-(t)\}$ $(i = 1, 2, 3, 4)$ to the eight vertices of a 4-dimensional cross-polytope (that is, to the standard basis vectors and their opposites).
- The projection to the last two coordinates $x_5, x_6$ sends the eight pairs of points $\{a_i^+(t), a_i^-(t)\}$ $(i = 1, \ldots, 8)$ to the eight vertices of a regular octagon.

The configuration $\mathcal{A}(0)$ already has a disconnected graph of triangulations.

**Theorem 3.2.** *There is a triangulation $K$ of the boundary of* $\mathrm{conv}(\mathcal{A}(0))$ *with the following two properties:*

1. *There are triangulations of $\mathcal{A}(0)$ inducing $K$ on the boundary.*
2. *No flip in a triangulation of $\mathcal{A}(0)$ inducing $K$ on the boundary affects the boundary.*

In fact, there are eight such triangulations. Hence:

[25] In oriented matroid language, the oriented matroid of $\mathcal{A}$ is a weak image of that of $\mathcal{A}'$.

**Corollary 3.3.** *The flip-graph of* $\mathcal{A}(0)$ *has at least nine connected components.*[26]

Of course, to describe the triangulation $K$ of the boundary of $\mathrm{conv}(\mathcal{A}_0)$ we need only specify how we triangulate each non-simplicial facet. The facets of $\mathrm{conv}(\mathcal{A}(0))$ are 96 simplices, and 16 non-simplicial facets $F_{\delta_1,\delta_2,\delta_3,\delta_4}$ ($\delta_i \in \{+,-\}$), each with eight vertices. More precisely,

$$F_{\delta_1,\delta_2,\delta_3,\delta_4} = \{a_1^{\delta_1}(0), a_2^{\delta_2}(0), a_3^{\delta_3}(0), a_4^{\delta_4}(0), a_5^{\delta_1}(0), a_6^{\delta_2}(0), a_7^{\delta_3}(0), a_8^{\delta_4}(0)\}.$$

All the $F_{*,*,*,*}$'s are equivalent under affine symmetries of $\mathcal{A}(0)$. For example, they are transitively permuted by the sixteen sign changes on the first four coordinates. Hence, the crucial point in the proof of Theorem 3.2 is to understand the triangulations of the point set $F_{+,+,+,+}$. This point set has dimension $d = 5$ and only eight ($= d+3$) points. In particular, all its triangulations are regular and their graph of flips is a cycle. Moreover, it is easy to check[27] that:

**Lemma 3.4.** 1. $\mathrm{conv}(F_{+,+,+,+})$ *has* 12 *facets. Eight of them are simplices and the other four have six points each, forming a* $(3,3)$ *circuit. In particular, there are sixteen ways to triangulate the boundary of* $F_{+,+,+,+}$.

2. $F_{+,+,+,+}$ *has eight triangulations.*

3. *Each flip in a triangulation of* $F_{+,+,+,+}$ *keeps the triangulation induced in three of the non-simplicial facets and switches the triangulation in the other.*

To construct the complex $K$ of Theorem 3.2 we choose the triangulations of the individual $F_{*,*,*,*}$ such that for every non-simplicial facet $G$ of an $F_{\delta_1,\delta_2,\delta_3,\delta_4}$, the triangulations chosen on $F_{\delta_1,\delta_2,\delta_3,\delta_4}$ and on the neighbor $F_{\delta_1',\delta_2',\delta_3',\delta_4'}$ agree on $G$ and one of them has the property that no flip on it changes the triangulation induced in $G$. In these conditions, no flip in any of the triangulations of the $F_{*,*,*,*}$'s is possible, since it would be incompatible with the triangulation of one of its neighbors.

**Example 3.5.** Lemma 3.4 implies, in particular, that only eight of the sixteen triangulations of the boundary of $F_{+,+,+,+}$ can be extended to the interior (without using additional interior points as vertices). Similar behavior occurs also in three-dimensional examples such as the set of vertices of a cube or a triangular prism.

Let us analyze the latter. It has three non-simplicial facets, whose vertex sets are $(2,2)$ circuits; in particular, there are eight ways to triangulate its boundary. But only six of them extend to the interior (all except the two "cyclic" ones). Each flip in a triangulation of $F_{+,+,+,+}$ keeps the triangulation induced in two of the non-simplicial facets and switches the triangulation in the other one.[28]

[26] Here, the ninth component is the one containing all the regular triangulations.

[27] For example, noting that a Gale transform of $F_{+,+,+,+}$ consists again of the eight vertices of a regular octagon, except in different order.

[28] The reader probably has noticed the similarities between this example and the configuration $F_{+,+,+,+}$. These similarities, and the fact that the constructions in [64] and [67] are ultimately based on glueing triangular prisms to one another, reflect the truth in (an instance of) Gian Carlo Rota's fifth lesson [62].

Let us now look at the perturbations $\mathcal{A}(t)$ and $\mathcal{A}'(t)$. The fact that $\mathcal{A}(t)$ (or $\mathcal{A}'(t)$) is a perturbation of $\mathcal{A}(0)$ implies that every triangulation of $\mathcal{A}(0)$ is still a geometric simplicial complex on $\mathcal{A}(t)$, except it may not cover the whole convex hull. In particular, the triangulation $K$ of the boundary of $\mathrm{conv}(\mathcal{A}_(0))$ mentioned in Theorem 3.2 can be embedded as a simplicial complex on $\mathcal{A}(t)$. We still call $K$ this perturbed simplicial complex. Then, Theorem 3.1 follows from the following more precise statement.

**Theorem 3.6.** *Let $t$ be a sufficiently small and positive constant. Then:*

1. *There are triangulations of $\mathcal{A}(t)$ containing the simplicial complex $K$.*
2. *If $T$ is a triangulation of $\mathcal{A}(t)$ containing the simplicial complex $K$, then every triangulation obtained from $T$ by a flip contains the simplicial complex $K$. In particular, the graph of triangulations of $\mathcal{A}(t)$ is not connected.*
3. *The previous two statements remain true if $\mathcal{A}(t)$ is perturbed into general position in an arbitrary way.*

## References


[1] Abramovich, D., Karu, K., Matsuki, K., Włodarczyk, J., Torification and factorization of birational maps. *J. Amer. Math. Soc.* **15** (3) (2002), 531–572.

[2] Abramovich, D., Matsuki, K., and Rashid, S., A note on the factorization theorem of toric birational maps after Morelli and its toroidal extension. *Tohoku Math. J.* (2) **51** (4) (1999), 489–537.

[3] Ambjorn, J., Carfora, M., Marzuoli, A., *The geometry of dynamical triangulations*. Lecture Notes in Phys. New Ser. m Monogr. 50, Springer-Verlag, Berlin 1997.

[4] Alexeev, V., Complete moduli in the presence of semiabelian group action. *Ann. of Math.* **155** (3) (2002), 611–708.

[5] Anderson, L., Matroid bundles and sphere bundles. In *New Perspectives in Algebraic Combinatorics* (ed. by L. J. Billera, A. Björner, C. Greene, R. E. Simion and R. P. Stanley), Math. Sci. Res. Inst. Publ. 38, Cambridge University Press, Cambridge 1999, 1–21.

[6] Azaola, M., The Baues conjecture in corank 3. *Topology* **41** (1) (2002), 183–209.

[7] Azaola, M., and Santos, F., The graph of triangulations of a point configuration with $d$+4 vertices is 3-connected. *Discrete Comput. Geom.* **23** (4) (2000), 489–536.

[8] Aurenhammer, F., Klein, R., Voronoi diagrams. In *Handbook of Computational Geometry* (ed. by J.-R. Sack and J. Urrutia), North-Holland, Amsterdam 2000, 201–290.

[9] Baues, H. J., Geometry of loop spaces and the cobar construction. *Mem. Amer. Math. Soc.* **25** (1980), 99–124.

[10] Bern, M., Triangulations and mesh generation. In *Handbook of Discrete and Computational Geometry* (ed. by J. E. Goodman and J. O'Rourke), 2nd edition, CRC Press Ser. Discrete Math. Appl., CRC Press, Boca Raton, FL, 2004, 563–582.

[11] Billera, L., Filliman, P., and Sturmfels, B., Constructions and complexity of secondary polytopes. *Adv. Math.* **83** (1990), 155–179.

[12] Billera, L., Gel'fand, I. M., and Sturmfels, B., Duality and minors of secondary polyhedra. *J. Combin. Theory Ser. B* **57** (2) (1993), 258–268.

[13] Billera, L., Kapranov, M. M., and Sturmfels, B., Cellular strings on polytopes. *Proc. Amer. Math. Soc.* **122** (2) (1994), 549–555.

[14] Billera, L., and Sturmfels, B., Fiber polytopes. *Ann. of Math.* **135** (1992), 527–549.

[15] Biss, D. K., The homotopy type of the matroid Grassmannian. *Ann. of Math.* (2) **158** (3) (2003), 929–952.

[16] Björner, A., Topological methods. In *Handbook of Combinatorics* (ed. by R. L. Graham, M. Grötschel and L. Lovász), Elsevier, Amsterdam 1995, 1819–1872.

[17] Björner, A., Las Vergnas, M., Sturmfels, B., White, N., and Ziegler, G. M., *Oriented Matroids*. 2nd edition, Cambridge University Press, Cambridge 1999.

[18] Björner, A., Lutz, F., Simplicial manifolds, bistellar flips and a 16-vertex triangulation of the Poincaré homology 3-sphere. *Experiment. Math.* **9** (2) (2000), 275–289.

[19] Clarkson, K. L., Shor, P. W., Applications of random sampling in computational geometry. II. *Discrete Comput. Geom.* **4** (5) (1989), 387–421.

[20] Connelly, R., Henderson, D. W., A convex 3-complex not simplicially isomorphic to a strictly convex complex. *Math. Proc. Cambridge Philos. Soc.* **88** (2) (1980), 299–306.

[21] de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O., *Computational geometry*. Algorithms and applications, second, revised edition, Springer-Verlag, Berlin 2000.

[22] de Loera, J. A., Rambau, J., and Santos, F., *Triangulations of polyhedra and point sets*. Book in preparation.

[23] de Loera, J. A., Santos, F., and Urrutia, J., The number of geometric bistellar neighbors of a triangulation. *Discrete Comput. Geom.* **21** (1) (1999), 131–142.

[24] Dougherty, R., Faber, V., and Murphy, M., Unflippable tetrahedral complexes. *Discrete Comput. Geom.* **32** (3) (2004), 309–315.

[25] Edelsbrunner, H., *Geometry and topology for mesh generation*. Cambridge Monogr. Appl. Comput. Math. 7, Cambridge University Press, Cambridge 2001.

[26] Edelsbrunner, H., and Shah, N. R., Incremental topological flipping works for regular triangulations. *Algorithmica* **15** (1996), 223–241.

[27] Fulton, W., *Introduction to toric varieties*. Ann. of Math. Stud., Princeton University Press, Princeton, NJ, 1993.

[28] Gel'fand, I. M., Kapranov, M. M., and Zelevinsky, A. V., Discriminants of polynomials in several variables and triangulations of Newton polyhedra. *Algebra i Analiz* **2** (3) (1990), 1–62; English transl. *Leningrad Math. J.* **2** (3) (1991), 449–505.

[29] Gel'fand, I. M., Kapranov, M. M., and Zelevinsky, A. V., *Discriminants, Resultants and Multidimensional Determinants*. Math. Theory Appl., Birkhäuser, Boston 1994.

[30] Gel'fand, I. M., and MacPherson, R. D., A combinatorial formula for the Pontrjagin classes. *Bull. Amer. Math. Soc.* (*N.S.*) **26** (2) (1992), 304–309.

[31] Grünbaum, B., *Convex polytopes*. Pure Appl. Math. 16, Interscience Publishers John Wiley & Sons, Inc., New York 1967; second edition, prepared and with a preface by V. Kaibel, V. Klee and G. M. Ziegler, Grad. Texts in Math. 221, Springer-Verlag, New York 2003.

[32] Guibas, L. J., Knuth, D. E., and Sharir, M., Randomized incremental construction of Delaunay and Voronoi diagrams. *Algorithmica* **7** (1992), 381–413.

[33] Haiman, M., and Sturmfels, B., Multigraded Hilbert schemes. *J. Algebraic Geom.* **13** (4) (2004), 725–769.

[34] Huber, B., Rambau, J., and Santos, F., The Cayley Trick, lifting subdivisions and the Bohne-Dress theorem on zonotopal tilings. *J. Eur. Math. Soc.* (*JEMS*) **2** (2) (2000), 179–198.

[35] Hurtado, F., Noy, M., Urrutia, J., Flipping edges in triangulations. *Discrete Comput. Geom.* **22** (1999), 333–346.

[36] Itenberg, I., Shustin, E., Viro theorem and topology of real and complex combinatorial hypersurfaces. *Israel J. Math.* **133** (2003), 189–238.

[37] Joe, B., Three dimensional triangulations from local transformations. *SIAM J. Sci. Statist. Comput.* **10** (1989), 718–741.

[38] Joe, B., Construction of three-dimensional Delaunay triangulations using local transformations. *Comput. Aided Geom. Design* **8** (1991), 123–142.

[39] Kapranov, M. M., Sturmfels, B., and Zelevinsky, A. V., Quotients of toric varieties. *Math. Ann.* **290** (1991), 643–655.

[40] Lawson, C. L., Transforming triangulations. *Discrete Math.* **3** (1972), 365–372.

[41] Lawson, C. L., Software for $C^1$ surface interpolation. In *Mathematical software. III*, Proceedings of a Symposium conducted by the Mathematics Research Center, the University of Wisconsin, Madison, Wis., 1977, Academic Press, New York 1977, 195–224.

[42] Lawson, C. L., Properties of $n$-dimensional triangulations. *Comput. Aided Geom. Design* **3** (4) (1986), 231–246.

[43] Lee, C. W., The associahedron and triangulations of the $n$-gon. *European J. Combin.* **10** (1990), 551–560.

[44] Lee, C. W., Regular triangulations of convex polytopes. In *Applied Geometry and Discrete Mathematics—The Victor Klee Festschrift* (ed. by P. Gritzmann and B. Sturmfels), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 4, Amer. Math. Soc., Providence, RI, 1991, 443–456.

[45] Lee, C. W., Subdivisions and triangulations of polytopes. In *Handbook of Discrete and Computational Geometry* (ed. by J. E. Goodman and J. O'Rourke), 2nd edition, CRC Press Ser. Discrete Math. Appl., CRC Press, Boca Raton, FL, 2004, 383–406.

[46] Lickorish, W. B. R., Simplicial moves on complexes and manifolds. In *Proceedings of the Kirbyfest* (Berkeley, CA, 1998), Geom. Topol. Monogr. 2, Geom. Topol. Publ., Coventry 1999, 299–320.

[47] Maclagan, D., and Thomas, R., Combinatorics of the toric Hilbert scheme. *Discrete Comput. Geom.* **27** (2002), 249–264.

[48] MacPherson, R. D., Combinatorial differential manifolds. In *Topological Methods in Modern Mathematics: a Symposium in Honor of John Milnor's Sixtieth Birthday* (ed. by L. R. Goldberg and A. Phillips), Publish or Perish, Houston 1993, 203–221.

[49] McMullen, P., Transforms, diagrams and representations. In *Contributions to geometry* (Proc. Geom. Sympos., Siegen, 1978), Birkhäuser, Basel 1979, 92–130.

[50] Mitchell, J. S. B., and O'Rourke, J., Computational geometry column 42. *Internat. J. Comput. Geom. Appl.* **11** (5) (2001), 573–582. Updated online as The Open Problems Project, http://maven.smith.edu/~orourke/TOPP/.

[51] Miyake, K., and Oda, T., *Torus embeddings and applications*. Tata Inst. Fund. Res. Lectures on Math. and Phys. 57, Springer-Verlag, Berlin 1978.

[52] Morelli, R., The birational geometry of toric varieties. *J. Alg. Geom.* **5** (1996), 751–782.

[53] Morelli, R., Correction to "The birational geometry of toric varieties". http://www.math.utah.edu/~morelli/Math/math.html.

[54] Nabutovsky, A., Geometry of the space of triangulations of a compact manifold. *Comm. Math. Phys.* **181** (2) (1996), 303–330.

[55] Oda, T., *Convex bodies and algebraic geometry*. Ergeb. Math. Grenzgeb. (3) 15, Springer-Verlag, Berlin 1988.

[56] Pachner, U., PL-homeomorphic manifolds are equivalent by elementary shellings. *European J. Combin.* **12** (1991), 129–145.

[57] Peeva, I., and Stillman, M., Toric Hilbert schemes. *Duke Math. J.* **111** (2002), 419–449.

[58] Radon, J., Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. *Math. Ann.* **83** (1921), 113–115.

[59] Rajan, V. T., Optimality of the Delaunay triangulation in $\mathbb{R}^d$. *Discrete Comput. Geom.* **12** (1994), 189–202.

[60] Rambau, J., and Ziegler, G. M., Projections of polytopes and the generalized Baues conjecture. *Discrete Comput. Geom.* **16** (1996), 215–237.

[61] Reiner, V., The generalized Baues problem. In *New Perspectives in Algebraic Combinatorics* (ed. by L. J. Billera, A. Björner, C. Greene, R. E. Simion and R. P. Stanley), Math. Sci. Res. Inst. Publ. 38, Cambridge University Press, Cambridge 1999, 293–336.

[62] Rota, G. C., Ten lessons I wish I had been taught. *Notices Amer. Math. Soc.* **44** (1) (1997), 22–25; reprinted in *Indiscrete thoughts*, Birkhäuser Boston, Inc., Boston, MA, 1997.

[63] Santos, F., Triangulations with very few geometric bistellar neighbors. *Discrete Comput. Geom.* **23** (2000), 15–33.

[64] Santos, F., A point configuration whose space of triangulations is disconnected. *J. Amer. Math. Soc.* **13** (3) (2000), 611–637.

[65] Santos, F., On the refinements of a polyhedral subdivision. *Collect. Math.* **52** (3) (2001), 231–256.

[66] Santos, F., Triangulations of Oriented Matroids. *Mem. Amer. Math. Soc.* **156** (741) (2002).

[67] Santos, F., Non-connected toric Hilbert schemes. *Math. Ann.* **332** (3) (2005), 645–665.

[68] Santos, F., A non-connected graph of triangulations in general position. Preprint in preparation.

[69] Schönhardt, E., Über die Zerlegung von Dreieckspolyedern in Tetraeder. *Math. Ann.* **98** (1928), 309–312.

[70] Shewchuk, J. R., Updating and Constructing Constrained Delaunay and Constrained Regular Triangulations by Flips. In *Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, ACM Press, New York 2003, 181–190.

[71] Sleator, D., Tarjan, R., and Thurston, W., Rotation distance, triangulations, and hyperbolic geometry. *J. Amer. Math. Soc.* **1** (1988), 647–681.

[72] Stasheff, J. D., Homotopy associativity of H-spaces. *Trans. Amer. Math. Soc.* **108** (1963), 275–292.

[73] Sturmfels, B., The geometry of $A$-graded algebras, Technical Report October 1994; http://www.arxiv.org/math.AG/9410032.

[74] Sturmfels, B., *Gröbner bases and convex polytopes*. University Series Lectures 8, Amer. Math. Soc., Providence, RI, 1995.

[75] Sturmfels, B., and Ziegler, G. M., Extension spaces of oriented matroids. *Discrete Comput. Geom.* **10** (1993), 23–45.

[76] Tamari, D., The algebra of bracketings and their enumeration. *Nieuw Arch. Wisk.* **10** (1962), 131–146.

[77] Viro, O. Ya., Gluing of plane real algebraic curves and constructions of curves of degrees 6 and 7. In *Topology* (Leningrad, 1982), Lecture Notes in Math. 1060, Springer-Verlag, Berlin 1984, 187–200.

[78] Włodarczyk, J., Decomposition of Birational Toric Maps in Blow-Ups and Blow-Downs. A Proof of the Weak Oda Conjecture. *Trans. Amer. Math. Soc.* **349** (1997), 373–411.

[79] Ziegler, G. M., *Lectures on polytopes*. Revised first edition, Grad. Texts in Math. 152, Springer-Verlag, New York, 1998.

[80] Ziegler, G. M., Recent progress on polytopes. In *Advances in Discrete and Computational Geometry* (ed. by B. Chazelle, J. E. Goodman, R. Pollack), Contemp. Math. 223, Amer. Math. Soc., Providence, RI, 1998, 395–406.

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, 39005 Santander, Spain

E-mail: francisco.santos@unican.es, URL: http://personales.unican.es/santosf/

# A survey of Pfaffian orientations of graphs

Robin Thomas*

**Abstract.** An orientation of a graph $G$ is Pfaffian if every even cycle $C$ such that $G\setminus V(C)$ has a perfect matching has an odd number of edges directed in either direction of the cycle. The significance of Pfaffian orientations is that if a graph has one, then the number of perfect matchings (a.k.a. the dimer problem) can be computed in polynomial time.

The question of which bipartite graphs have Pfaffian orientations is equivalent to many other problems of interest, such as a permanent problem of Pólya, the even directed cycle problem, or the sign-nonsingular matrix problem for square matrices. These problems are now reasonably well-understood. On the other hand, it is not known how to efficiently test if a general graph is Pfaffian, but there are some interesting connections with crossing numbers and signs of edge-colorings of regular graphs.

**Mathematics Subject Classification (2000).** Primary 05C75; Secondary 05C10, 05C20, 05C38, 05C70, 05C83, 05C85, 68R10, 82B20.

**Keywords.** Graph, matching, dimer problem, Pfaffian orientation, even directed cycle, brick, brace, Pólya's permanent problem, sign-nonsingular matrix, crossing number.

## 1. Introduction

All *graphs* in this paper are finite, do not have loops or multiple edges and are undirected. *Directed graphs*, or *digraphs*, do not have loops or multiple edges, but may have two edges between the same pair of vertices, one in each direction. Most of our terminology is standard and can be found in many textbooks, such as [4], [10], [65]. In particular, *cycles* and *paths* have no repeated vertices. A subgraph $H$ of a graph $G$ is called *central* if $G\setminus V(H)$ has a perfect matching (we use $\setminus$ for deletion). An even cycle $C$ in a directed graph $D$ is called *oddly oriented* if for either choice of direction of traversal around $C$, the number of edges of $C$ directed in the direction of traversal is odd. Since $C$ is even, this is clearly independent of the initial choice of direction of traversal. Finally, an orientation $D$ of (the edges of) a graph $G$ is *Pfaffian* if every even central cycle of $G$ is oddly oriented in $D$. We say that a graph $G$ is *Pfaffian* if it has a Pfaffian orientation.

The significance of Pfaffian orientations stems from the fact that if a graph $G$ has one, then the number of perfect matchings of $G$ (as well as other related problems) can be computed in polynomial time. We survey this in Section 2. The following

*Partially supported by NSF grants DMS-0200595 and 0354742.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

is a classical theorem of Kasteleyn [23]. A special case is implicit in the work of Fisher [16] and Temperley and Fisher [55]. Different proofs may be found in [25], [26], [32], [38].

**Theorem 1.1.** *Every planar graph is Pfaffian.*

The smallest non-Pfaffian graph is the complete bipartite graph $K_{3,3}$. This paper is centered around the question of which graphs are Pfaffian. For bipartite graphs this is equivalent to many other problems of interest, and is by now reasonably well-understood. We list several such problems in Section 3, including a question of Pólya from 1913 whether the permanent of a square matrix can be calculated by a reduction to a determinant of a related matrix, the even directed cycle problem for digraphs, and the sign-nonsingular matrix problem. In Section 4 we discuss two characterizations of bipartite Pfaffian graphs. The first is in terms of excluded obstructions; it turns out that for bipartite graphs $K_{3,3}$ is the only obstruction with respect to the "matching minor" partial order, defined later. This is an analogue of the graph minor relation and is well-suited for problems involving perfect matchings. Unfortunately, it no longer has many of the nice properties of the usual minor order. The second characterization is structural and describes the structure of all bipartite Pfaffian graphs. It turns out those graphs and only those graphs can be built from planar graphs and one sporadic nonplanar graph by certain composition operations. This characterization implies a polynomial-time algorithm to decide whether a bipartite graph is Pfaffian, and hence solves all the problems listed in Section 3. Applications of the structure theorem are discussed in Section 5.

We then turn to general graphs. In Section 6 we review a matching decomposition procedure of Lovász and Plummer that decomposes every graph into "bricks" and "braces". The decomposition has the property that a graph is Pfaffian if and only if all its constituent bricks and braces are Pfaffian. Furthermore, braces are bipartite, and hence whether they are Pfaffian can be decided using the algorithm of Section 4. Thus in order to test whether an input graph is Pfaffian it suffices to design an algorithm for bricks. Motivated by this we present a recent theorem that describes how to construct an arbitrary brick, and later we discuss various examples and results that were obtained using this theorem.

In the next section we talk about results of Norine that relate Pfaffian graphs and crossing numbers. The starting point here is Theorem 7.1 that characterizes Pfaffian graphs in terms of drawings in the plane. Norine then generalized it to $T$-joins, whereby the generalization implies several well-known results about crossing numbers, and in a different direction proved an analogue for 4-Pfaffian graphs and drawings in the torus. The latter suggests a general conjecture that is still open.

In Section 8 we discuss the relationship between signs of edge-colorings (in the sense of Penrose [46]) and Pfaffian orientations. We mention a proof of a conjecture of Goddyn that in a $k$-regular Pfaffian graph all $k$-edge-colorings have the same sign, which holds more generally for graphs that admit a "Pfaffian labeling." We present a partial converse of this, and then describe two characterizations of graphs that admit a

Pfaffian labeling. The above research led Norine and the author to make the following conjecture [44].

**Conjecture 1.2.** Every 2-connected 3-regular Pfaffian graph is 3-edge-colorable.

Let us recall that by Tait's result [54] (see also [65]) the Four-Color Theorem is equivalent to the statement that every 2-connected 3-regular planar graph is 3-edge-colorable. Thus, if true, Conjecture 1.2 would imply the Four-Color Theorem by Theorem 1.1.

In the last section we discuss the prospects for characterizing general Pfaffian graphs, either structurally or by means of excluded matching minors.

## 2. Pfaffian orientations and counting perfect matchings

Pfaffian orientations were invented by the physicists M. E. Fisher, P. W. Kasteleyn, and H. N. V. Temperley as a tool for enumerating the number of perfect matchings in a graph (or, in physics terminology, to solve the dimer problem). Let us start by explaining their approach. Let $A = (a_{ij})$ be a skew symmetric $n \times n$ matrix; that is $a_{ij} = -a_{ji}$. For each partition $\pi = \{\{i_1, j_1\}, \{i_2, j_2\}, \ldots, \{i_k, j_k\}\}$ of the set $\{1, 2, \ldots, n\}$ into unordered pairs ("partition into pairs") we define the quantity

$$\sigma_\pi = \operatorname{sgn}\begin{pmatrix} 1 & 2 & 3 & 4 & \ldots & 2k-1 & 2k \\ i_1 & j_1 & i_2 & j_2 & \ldots & i_k & j_k \end{pmatrix} a_{i_1 j_1} a_{i_2 j_2} \ldots a_{i_k j_k}, \tag{1}$$

where sgn denotes the sign of the indicated permutation. Clearly, there is no partition into pairs if $n$ is odd. The *Pfaffian* of $A$ is defined by $\operatorname{Pf}(A) = \sum \sigma_\pi$, where the summation is over all partitions of $\{1, 2, \ldots, n\}$ into pairs. Since $A$ is skew symmetric the value of $\sigma_\pi$ does not depend on the order of blocks of $\pi$ or on the order in which the members of a block are listed, and hence $\operatorname{Pf}(A)$ is well-defined. We will need the following lemma from linear algebra [23], [37].

**Lemma 2.1.** *If $A$ is a skew symmetric matrix, then* $\det A = (\operatorname{Pf}(A))^2$.

Now let $G$ be a graph with vertex-set $\{1, 2, \ldots, n\}$, and let $D$ be an orientation of (the edges of) $G$. To the orientation $D$ there corresponds a skew adjacency matrix $A = (a_{ij})$ of $G$ defined by saying that $a_{ij} = 0$ if $i$ is not adjacent to $j$, and otherwise $a_{ij} = 1$ if the edge $ij$ is directed in $D$ from $i$ to $j$ and $a_{ij} = -1$ if the edge $ij$ is directed in $D$ from $j$ to $i$. If $\pi$ is a partition of $\{1, 2, \ldots, n\}$ into pairs, then $\sigma_\pi \neq 0$ if and only if each pair in $\pi$ is an edge of $G$, or, in other words, $\pi$ is a perfect matching of $G$. Thus the summation in the definition of $\operatorname{Pf}(A)$ might as well be restricted to perfect matchings of $G$. We define $\operatorname{sgn}_D(M)$, the *sign of a perfect matching $M$ of $D$*, as $\sigma_M$, or, equivalently, by

$$\operatorname{sgn}_D(M) = \operatorname{sgn}\begin{pmatrix} 1 & 2 & 3 & 4 & \ldots & 2k-1 & 2k \\ i_1 & j_1 & i_2 & j_2 & \ldots & i_k & j_k \end{pmatrix}, \tag{2}$$

where the edges of $M$ are listed as $i_1j_1, i_2j_2, \ldots, i_kj_k$ in such a way that $i_tj_t$ is directed from $i_t$ to $j_t$ in $D$. It is not hard to see that $D$ is a Pfaffian orientation of $G$ if and only if $\mathrm{sgn}_D(M)$ does not depend on $M$. If that is the case, then $|\mathrm{Pf}(A)|$ is equal to the number of perfect matchings of $G$, and by Lemma 2.1 the number of perfect matchings of $G$ can be computed efficiently.

This is significant, because Valiant [63] proved that counting the number of perfect matchings in general graphs (even in bipartite graphs) is #P-complete, and therefore is unlikely to be polynomial-time solvable. Furthermore, Theorem 1.1 guarantees that there is an interesting and useful class of graphs for which this technique can be applied.

The dimer problem of statistical mechanics is concerned with the properties of a system of diatomic molecules, or dimers, adsorbed on the surface of a crystal. Usually it is assumed that the adsorption points form the vertices of a lattice graph, such as the 2-dimensional grid. A crucial problem in the calculation of the thermodynamic properties of such a system of dimers is that of enumerating all ways in which a given number of dimers can be arranged on the lattice without overlapping each other. In the related monomer-dimer model some sites may be left unoccupied, but in the dimer model it is assumed that the dimers cover all the vertices of the graph; in other words, they form a perfect matching. Kasteleyn [21], [22], [23], Fisher [16] and Temperley and Fisher [55] used the method described in this section to solve the 2-dimensional dimer problem. The method is more general in the sense that it allows the computation of the dimer partition function, and that, in turn, can be used to solve the 2-dimensional Ising problem [23]. Let us remark that the 3-dimensional dimer problem remains open.

## 3. Some equivalent problems

Vazirani and Yannakakis [64] used a deep theorem of Lovász [31] to show the following.

**Theorem 3.1.** *The decision problems "Is a given orientation of a graph Pfaffian" and "Is an input graph Pfaffian" are polynomial-time equivalent.*

This is reasonably easy for bipartite graphs. There does not seem to be an elementary proof for general graphs, but the theorem can be easily deduced from the results discussed in Section 6.

Computing the permanent of a matrix seems to be of a different computational complexity from computing the determinant. While the determinant can be calculated using Gaussian elimination, no efficient algorithm for computing the permanent is known, and, in fact, none is believed to exist. More precisely, Valiant [63] has shown that computing the permanent is #P-complete even when restricted to 0-1 matrices.

It is therefore reasonable to ask if perhaps computing the permanent can be somehow reduced to computing the determinant of a related matrix. In particular, the

following question was asked by Pólya [47] in 1913. If $A$ is a 0-1 square matrix, does there exist a matrix $B$ obtained from $A$ by changing some of the 1's to $-1$'s in such a way that the permanent of $A$ equals the determinant of $B$? For the purpose of this paper let us say that $B$ (when it exists) is a *Pólya matrix* for $A$.

Let $G$ be a bipartite graph with bipartition $(X, Y)$. The *bipartite adjacency matrix of* $G$ has rows indexed by $X$, columns indexed by $Y$, and the entry in row $x$ and column $y$ is 1 or 0 depending on whether $x$ is adjacent to $y$ or not. Vazirani and Yannakakis [64] proved the following.

**Theorem 3.2.** *Let $G$ be a bipartite graph, and let $A$ be its bipartite adjacency matrix. Then $A$ has a Pólya matrix if and only if $G$ has a Pfaffian orientation.*

Let us turn to directed graphs now. A digraph $D$ is *even* if for every weight function $w\colon E(D) \to \{0, 1\}$ there exists a cycle in $D$ of even total weight. It was shown in [53] and is not difficult to see that testing evenness is polynomial-time equivalent to testing whether a digraph has an even directed cycle. (This is equivalent to Theorem 3.1 for bipartite graphs.) Let $G$ be a bipartite graph with bipartition $(A, B)$, and let $M$ be a perfect matching in $G$. Let $D = D(G, M)$ be obtained from $G$ by directing every edge from $A$ to $B$, and contracting every edge of $M$. Little [27] has shown the following.

**Lemma 3.3.** *Let $G$ be a bipartite graph, and let $M$ be a perfect matching in $G$. Then $G$ has a Pfaffian orientation if and only if $D(G, M)$ is not even.*

We say that two $n \times m$ matrices $A = (a_{ij})$ and $B = (b_{ij})$ have the same *sign-pattern* if for all pairs of indices $i$, $j$ the entries $a_{ij}$ and $b_{ij}$ have the same sign; that is, they are both strictly positive, or they are both strictly negative, or they are both zero. A square matrix $A$ is *sign-nonsingular* if every real matrix with the same sign pattern is nonsingular.

In economic analysis one may not know the exact quantitative relationships between different variables, but there may be some qualitative information such as that one quantity rises if and only if another does. For instance, it is generally agreed that the supply of a particular commodity increases as the price increases, even though the exact dependence may vary. Thus we may want to deduce qualitative information about the solution to a linear system $A\boldsymbol{x} = \boldsymbol{b}$ from the knowledge of the sign-patterns of the matrix $A$ and vector $\boldsymbol{b}$. That motivates the following definition. We say that the linear system $A\boldsymbol{x} = \boldsymbol{b}$ is *sign-solvable* if for every real matrix $B$ with the same sign-pattern as $A$ and every vector $\boldsymbol{c}$ with the same sign-pattern as $\boldsymbol{b}$ the system $B\boldsymbol{y} = \boldsymbol{c}$ has a unique solution $\boldsymbol{y}$, and its sign-pattern does not depend on the choice of $B$ and $\boldsymbol{c}$. The study of sign-solvability was first proposed by Samuelson [51].

It follows from standard linear algebra that sign-solvability can be decided efficiently if and only if sign-nonsingularity can. But for square matrices the latter is equivalent to testing whether a given orientation of a bipartite graph is Pfaffian. To state the result, let $D$ be a bipartite digraph with bipartition $(X, Y)$. The *directed bipartite adjacency matrix of* $D$ has rows indexed by $X$, columns indexed by $Y$, and the entry in row $x$ and column $y$ is 1, $-1$ or 0 depending on whether $D$ has an edge

directed from $x$ to $y$, or $D$ has an edge directed from $y$ to $x$, or $x$ and $y$ are not adjacent in $D$. By Theorem 3.1 the following result implies that testing sign-solvability is polynomial-time equivalent to testing whether a bipartite graph is Pfaffian.

**Theorem 3.4.** *Let $D$ be a directed bipartite graph with a perfect matching, and let $A$ be its directed bipartite adjacency matrix. Then $A$ is sign-nonsingular if and only if $D$ is a Pfaffian orientation of its underlying undirected graph.*

The next problem is about hypergraph coloring. A *hypergraph* $H$ is a pair $(V(H), E(H))$, where $V(H)$ is a finite set and $E(H)$ is a collection of distinct nonempty subsets of $V(H)$. We say that $H$ is 2-*colorable* if $V(H)$ can be colored using two colors in such a way that every edge includes vertices of both colors. We say that $H$ is *minimally non*-2-*colorable* if $H$ is not 2-colorable, has no isolated vertices, and the deletion of any member of $E(H)$ results in a 2-colorable hypergraph. Seymour [52] proved the following.

**Theorem 3.5.** *Let $H$ be a hypergraph with no isolated vertices and $|E(H)| = |V(H)|$, let $D$ be the digraph with bipartition $(V(H), E(H))$ defined by saying that $D$ has an edge directed from $v \in V(H)$ to $E \in E(H)$ if and only if $v \in E$, and let $G$ be the underlying undirected graph of $D$. Then $H$ is minimally non-2-colorable if and only if $G$ is connected, every edge of $G$ belongs to a perfect matching of $G$ and $D$ is a Pfaffian orientation of $G$.*

Our last problem is about the polytope of even permutation matrices. The convex hull of permutation matrices has been characterized by Birkhoff [3] as precisely the set of doubly stochastic matrices. It is an open problem to characterize the convex hull of even permutation matrices. More precisely, it is not known if there exists a polynomial-time algorithm to test whether a given $n \times n$ matrix belongs to this polytope. By a fundamental result of Grötschel, Lovász and Schrijver [19] this problem is solvable in polynomial time if there exists a polynomial-time algorithm for the optimization problem: Given a fixed $n \times n$ matrix $M$, find the maximum of $M \cdot X$ over all even permutation matrices $X$, where "$\cdot$" denotes the dot product in $\mathbb{R}^{n^2}$ and both matrices are regarded as vectors of length $n^2$.

A special case of the above optimization problem when $A$ is a 0-1 matrix and we want to determine if the maximum is $n$ can be reformulated as follows. Let $G$ be a bipartite graph with bipartition $(A, B)$, and let $D$ be the orientation of $G$ defined by orienting every edge from $A$ to $B$. The problem is: "Decide if $G$ has a perfect matching $M$ such that $\mathrm{sgn}_D(M) = 1$." By Theorem 3.1 this is polynomial-time equivalent to deciding whether a bipartite graph has a Pfaffian orientation.

## 4. Characterizing bipartite Pfaffian graphs

We have seen in the previous section that characterizing bipartite Pfaffian graphs is of interest. In this section we discuss two such characterizations and a recognition

algorithm. We begin with an elegant theorem of Little [27]. Let $H$ be a graph, and let $v$ be a vertex of $H$ of degree two. By *bicontracting* $v$ we mean contracting both edges incident with $v$ and deleting the resulting loops and parallel edges. A graph $G$ is a *matching minor* of a graph $H$ if $G$ can be obtained from a central subgraph of $H$ by repeatedly bicontracting vertices of degree two. It is fairly easy to see that a matching minor of a Pfaffian graph is Pfaffian.

**Theorem 4.1.** *A bipartite graph admits a Pfaffian orientation if and only if it has no matching minor isomorphic to* $K_{3,3}$.

By Lemma 3.3 the above implies a characterization of even digraphs. Seymour and Thomassen obtained such characterization from first principles in [53]. Interestingly, the latter involves infinitely many excluded minors, rather than one.

Unfortunately, Theorem 4.1 does not seem to imply a polynomial-time algorithm to test whether a bipartite graph is Pfaffian, the difficulty being that it is not clear how to efficiently test for the presence of a matching minor isomorphic to $K_{3,3}$. The next result gives a structural description of bipartite Pfaffian graphs, and can be used to derive a polynomial-time recognition algorithm. We need some definitions first.

Let $G_0$ be a graph, let $C$ be a central cycle of $G_0$ of length four, and let $G_1, G_2, G_3$ be three subgraphs of $G_0$ such that $G_1 \cup G_2 \cup G_3 = G_0$, and for distinct integers $i, j \in \{1, 2, 3\}$, $G_i \cap G_j = C$ and $V(G_i) - V(C) \neq \emptyset$. Let $G$ be obtained from $G_0$ by deleting some (possibly none) of the edges of $C$. In these circumstances we say that $G$ is a *trisum* of $G_1$, $G_2$ and $G_3$. The *Heawood graph* is the bipartite graph associated with the incidence matrix of the Fano plane (see Figure 1).

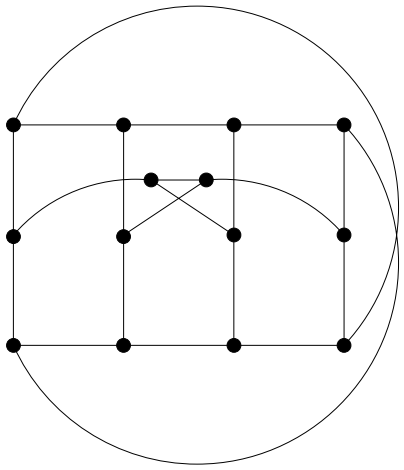


Figure 1. The Heawood graph.

A graph $G$ is *k-extendable*, where $k \geq 0$ is an integer, if every matching of size at most $k$ can be extended to a perfect matching. A connected 2-extendable bipartite

graph is called a *brace*. It is easy to see (and will be outlined in Section 6) that the problem of finding Pfaffian orientations of bipartite graphs can be reduced to braces. The following was shown in [35] and, independently, in [50].

**Theorem 4.2.** *A brace has a Pfaffian orientation if and only if either it is isomorphic to the Heawood graph, or it can be obtained from planar braces by repeated application of the trisum operation.*

Let us turn to testing whether a bipartite graph is Pfaffian. We wish to apply Theorem 4.2, and for that the following result [50, Theorem 8.3] is very helpful.

**Theorem 4.3.** *Let $G$ be a brace that has a Pfaffian orientation, and let $G$ be a trisum of $G_1$, $G_2$ and $G_3$. Then $G_1$, $G_2$ and $G_3$ have a Pfaffian orientation.*

A polynomial-time algorithm now follows easily. Given a bipartite graph $G$ we first decompose it into braces (more on that in Section 6), and apply the algorithm recursively to each brace in the decomposition. Thus we may assume that $G$ is a brace. Now we test if $G$ has a set $X \subseteq V(G)$ of size four such that $G\backslash X$ has at least three components. If it does, then $G$ can be expressed as a trisum of three smaller graphs, and by Theorem 4.3 we may apply the algorithm recursively to each of the three smaller graphs. On the other hand, if $G$ has no set $X$ as above, then by Theorem 4.2 $G$ is Pfaffian if and only if it is planar or isomorphic to the Heawood graph. It is clear that this is a polynomial-time algorithm. In [50] it is shown how to implement it to run in time $O(|V(G)|^3)$. By using more modern algorithmic results the running time can be reduced to $O(|V(G)|^2)$.

## 5. Applications of the characterization of bipartite Pfaffian graphs

As a corollary of Theorem 4.2 we get the following extremal result.

**Corollary 5.1.** *No brace with $n \geq 3$ vertices and more than $2n - 4$ edges has a Pfaffian orientation.*

*Proof.* Every planar bipartite graph on $n \geq 3$ vertices has at most $2n - 4$ edges. The result follows from Theorem 4.2 by induction. □

Since every digraph is isomorphic to $D(G, M)$ for some $G$ and $M$, Theorem 4.2 gives a characterization of even directed graphs, using Lemma 3.3. Let us state the characterization explicitly, but first let us point out a relation between extendability and strong connectivity. A digraph $D$ is *strongly connected* if for every two vertices $u$ and $v$ it has a directed path from $u$ to $v$. It is *strongly $k$-connected*, where $k \geq 1$ is an integer, if for every set $X \subseteq V(D)$ of size less than $k$, the digraph $D\backslash X$ is strongly connected. The following is straightforward.

**Lemma 5.2.** *Let $G$ be a connected bipartite graph, let $M$ be a perfect matching in $G$, and let $k \geq 1$ be an integer. Then $G$ is $k$-extendable if and only if $D(G, M)$ is strongly $k$-connected.*

Let $D$ be a digraph, and let $(X, Y)$ be a partition of $V(G)$ into two nonempty sets in such a way that no edge of $G$ has tail in $X$ and head in $Y$. Let $D_1 = D\backslash Y$ and $D_2 = D\backslash X$. We say that $D$ is a 0-*sum* of $D_1$ and $D_2$. Now let $v \in V(D)$, and let $(X, Y)$ be a partition of $V(D) - \{v\}$ into two nonempty sets such that no edge of $D$ has tail in $X$ and head in $Y$. Let $D_1$ be obtained from $D$ by deleting all edges with both ends in $Y \cup \{v\}$ and identifying all vertices of $Y \cup \{v\}$, and let $D_2$ be obtained by deleting all edges with both ends in $X \cup \{v\}$ and identifying all vertices of $X \cup \{v\}$. We say that $D$ is a 1-*sum* of $D_1$ and $D_2$. Let $D_0$ be a directed graph, let $u, v \in V(D_0)$, and let $uv, vu \in E(D_0)$. Let $D_1$ and $D_2$ be such that $D_1 \cup D_2 = D_0$, $V(D_1) \cap V(D_2) = \{u, v\}$, $V(D_1) - V(D_2) \neq \emptyset \neq V(D_2) - V(D_1)$ and $E(D_1) \cap E(D_2) = \{uv, vu\}$. Let $D$ be obtained from $D_0$ by deleting some (possibly neither) of the edges $uv, vu$. We say that $D$ is a 2-*sum* of $D_1$ and $D_2$. Now let $D_0$ be a directed graph, let $u, v, w \in V(D_0)$, let $uv, wv, wu \in E(D_0)$, and assume that $D_0$ has a directed cycle containing the edge $wv$, but not the vertex $u$. Let $D_1$ and $D_2'$ be such that $D_1 \cup D_2' = D_0$, $V(D_1) \cap V(D_2') = \{u, v, w\}$, $V(D_1) - V(D_2') \neq \emptyset \neq V(D_2') - V(D_1)$ and $E(D_1) \cap E(D_2') = \{uv, wv, wu\}$, let $D_2'$ have no edge with tail $v$, and no edge with head $w$. Let $D$ be obtained from $D_0$ by deleting some (possibly none) of the edges $uv, wv, wu$, and let $D_2$ be obtained from $D_2'$ by contracting the edge $wv$. We say that $D$ is a 3-*sum* of $D_1$ and $D_2$. Finally let $D_0$ be a directed graph, let $x, y, u, v \in V(D_0)$, let $xy, xv, uy, uv \in E(D_0)$, and assume that $D_0$ has a directed cycle containing precisely two of the edges $xy, xv, uy, uv$. Let $D_1$ and $D_2'$ be such that $D_1 \cup D_2' = D_0$, $V(D_1) \cap V(D_2') = \{x, y, u, v\}$, $V(D_1) - V(D_2') \neq \emptyset \neq V(D_2') - V(D_1)$ and $E(D_1) \cap E(D_2') = \{xy, xv, uy, uv\}$, let $D_2'$ have no edge with tail $y$ or $v$, and no edge with head $x$ or $u$. Let $D$ be obtained from $D_0$ by deleting some (possibly none) of the edges $xy, xv, uy, uv$, and let $D_2$ be obtained from $D_2'$ by contracting the edges $xy$ and $uv$. We say that $D$ is a 4-*sum* of $D_1$ and $D_2$. We say that a digraph is *strongly planar* if it has a planar drawing such that for every vertex $v \in V(D)$, the edges of $D$ with head $v$ form an interval in the cyclic ordering of edges incident with $v$ determined by the planar drawing. Let $F_7$ be the directed graph $D(H, M)$, where $H$ is the Heawood graph, and $M$ is a perfect matching of $H$. This defines $F_7$ uniquely up to isomorphism, irrespective of the choice of the bipartition of $H$ or the choice of $M$. Lemma 3.3 and Theorem 4.2 imply the following.

**Theorem 5.3.** *A digraph $D$ is not even if and only if it can be obtained from strongly planar digraphs and $F_7$ by means of* 0-, 1-, 2-, 3- *and* 4-*sums.*

From Corollary 5.1 and Lemmas 3.3 and 5.2 we deduce the following extremal result.

**Corollary 5.4.** *Let $D$ be a strongly 2-connected directed graph on $n \geq 2$ vertices. If $D$ has more than $3n - 4$ edges, then $D$ is even.*

Corollary 5.4 does not hold for strongly connected digraphs. However, Thomassen [59] has shown that every strongly connected directed graph with minimum in- and out-degree at least three is even. This is equivalent to the following by Lemma 3.3.

**Corollary 5.5.** *Let $G$ be a 1-extendable bipartite graph such that every vertex has degree at least four. Then $G$ does not have Pfaffian orientation.*

If $G$ is a brace, then the corollary follows from Corollary 5.1; otherwise the corollary follows by induction using the matching decomposition explained in the next section. The details may be found in [50, Corollary 7.8].

In [33] McCuaig used Theorem 4.2 to answer a question of Thomassen [58] by proving the following.

**Theorem 5.6.** *The digraph $F_7$ is the unique strongly 2-connected digraph with no even cycle.*

## 6. Matching decomposition

We have seen in the preceding sections that the problem of understanding which bipartite graphs are Pfaffian is reasonably well-understood and has applications outside of this subfield. We now turn our attention to the same question for general graphs. This problem seems much harder, but there are some interesting and unexpected connections.

The brick decomposition procedure of Lovász and Plummer [32] can be used to reduce the question of characterizing Pfaffian graphs to "bricks". The purpose of this section is to give an overview of this decomposition technique and to discuss recent additions to it.

A graph is *matching covered* if it is connected and every edge belongs to a perfect matching. Clearly, when deciding whether a graph $G$ is Pfaffian we may assume that $G$ is matching covered, for edges that belong to no perfect matching may be deleted without affecting the outcome.

Let $G$ be a graph, and let $X \subseteq V(G)$. We use $\delta(X)$ to denote the set of edges with one end in $X$ and the other in $V(G) - X$. A *cut* in $G$ is any set of the form $\delta(X)$ for some $X \subseteq V(G)$. A cut $C$ is *tight* if $|C \cap M| = 1$ for every perfect matching $M$ in $G$. Every cut of the form $\delta(\{v\})$ in a graph with a perfect matching is tight; those are called *trivial*, and all other tight cuts are called *nontrivial*.

Here are three important examples of tight cuts. Let $G$ be a matching covered graph. Assume first that $G$ is bipartite with bipartition $(A, B)$, and that $G$ is not a brace. Then by Hall's theorem there is a set $X \subseteq A$ such that $|N(X)| = |X| + 1$ and $N(X) \neq B$, where $N(X)$ denotes the set of all vertices $v \in V(G) - X$ with a neighbor in $X$. Then $\delta(X \cup N(X))$ is a nontrivial tight cut. Now assume that $G$ is not bipartite. If there exist distinct vertices $u, v \in V(G)$ such that $G \backslash u \backslash v$ has no perfect matching, then by Tutte's 1-factor theorem [61] there exists a nonempty set $X \subseteq V(G)$ such

that $G\backslash X$ has exactly $|X|$ odd components. Furthermore, by repeatedly adding to $X$ one vertex from each even component of $G\backslash X$ we may assume that $G\backslash X$ has no even components. Since $G$ is matching covered no edge of $G$ has both ends in $X$, and since $G$ is not bipartite some component of $G\backslash X$, say $C$, has more than one vertex. But then $\delta(V(C))$ is a nontrivial tight cut. Finally, if $G$ is not 3-connected, then let $u$, $v$ be distinct vertices of $G$ such that $G\backslash u\backslash v$ is disconnected. Let $A$ be the vertex-set of a component of $G\backslash u\backslash v$ and let $B$ be the union of all the remaining components. Notice that if $|A|$ is odd, then $G\backslash u\backslash v$ has no perfect matching. If $|A|$ is even, then $\delta(A\cup\{u\})$ is a nontrivial tight cut.

It is not true that every nontrivial tight cut arises as described above, but Theorem 6.1 below implies that if a graph has a nontrivial tight cut, then it has a nontrivial tight cut that arises in one of the ways described in the previous paragraph. A *brick* is a 3-connected graph $G$ such that $G\backslash u\backslash v$ has a perfect matching for every two distinct vertices $u$, $v$ of $G$.

Let $\delta(X)$ be a nontrivial tight cut in a graph $G$, let $G_1$ be obtained from $G$ by identifying all vertices in $X$ into a single vertex and deleting all resulting parallel edges, and let $G_2$ be defined analogously by identifying all vertices in $V(G)-X$. Then many matching-related problems can be solved for $G$ if we are given the corresponding solutions for $G_1$ and $G_2$.

The above decomposition process can be iterated, until we arrive at graphs with no nontrivial tight cuts. Lovász [31] proved that the list of indecomposable graphs obtained at the end of the procedure does not depend on the choice of tight cuts made during the process. These indecomposable graphs were characterized by Edmonds, Lovász and Pulleyblank [12], [13]:

**Theorem 6.1.** *Let $G$ be a matching covered graph. Then $G$ has no nontrivial tight cut if and only if $G$ is a brick or a brace.*

In light of this theorem and the previous discussion we say that a brick or a brace $H$ is a *brick* or a *brace of a graph $G$* if $H$ is obtained when the tight cut decomposition procedure is applied to $G$.

Vazirani and Yannakakis [64] used the tight cut decomposition procedure to reduce the study of Pfaffian graphs to bricks and braces:

**Theorem 6.2.** *A graph $G$ is Pfaffian if and only if every brick and brace of $G$ is Pfaffian.*

In particular, this justifies our earlier claim that in order to understand Pfaffian bipartite graphs it suffices to understand Pfaffian braces. Since Pfaffian braces are characterized by Theorem 4.2, in order to understand Pfaffian graphs it suffices to understand Pfaffian bricks. We return to this problem later, but in the remainder of this section we describe a characterization of bricks, developed for the purpose of studying Pfaffian bricks. We need a few definitions first.

Let $G$ be a graph, and let $v_0$ be a vertex of $G$ of degree two incident with the edges $e_1 = v_0v_1$ and $e_2 = v_0v_2$. Let $H$ be obtained from $G$ by contracting both $e_1$

and $e_2$ and deleting all resulting parallel edges. We say that $H$ was obtained from $G$ by *bicontracting* or *bicontracting the vertex* $v_0$, and write $H = G/v_0$. Let us say that a graph $H$ is a *reduction* of a graph $G$ if $H$ can be obtained from $G$ by deleting an edge and bicontracting all resulting vertices of degree two. By a *prism* we mean the unique 3-regular planar graph on six vertices. The following is a generation theorem of de Carvalho, Lucchesi and Murty [9].

**Theorem 6.3.** *If $G$ is a brick other than $K_4$, the prism, and the Petersen graph, then some reduction of $G$ is a brick other than the Petersen graph.*

Thus if a brick $G$ is not the Petersen graph, then the reduction operation can be repeated until we reach $K_4$ or the prism. By reversing the process Theorem 6.3 can be viewed as a generation theorem.

Theorem 6.3 has interesting applications. First of all, it implies several results about various spaces generated by perfect matchings, including a deep theorem of Lovász [31] that characterizes the matching lattice of a graph. Second, it implies Theorem 3.1 (more precisely the most difficult part of that theorem, namely that it holds for bricks). Third, it can be used to prove a uniqueness theorem for Pfaffian orientations [8]:

**Theorem 6.4.** *A Pfaffian orientation of a graph $G$ can be transformed to any other Pfaffian orientation of $G$ by repeatedly applying the following operations:*

(1) *reversing the direction of all edges of a cut of $G$,*
(2) *reversing all edges with both ends in $S$ for some tight cut $\delta(S)$,*
(3) *reversing the direction of all edges of $G$.*

There is a strengthening of Theorem 6.3, which we now describe. First, the starting graph can be any matching minor of $G$ except $K_4$ and the prism, and second, reduction can be replaced by a more restricted operation, the following. We say that a graph $H$ is a *proper reduction* of a graph $G$ if it is a reduction in such a way that the bicontractions involved do not produce parallel edges. Unfortunately, Theorem 6.3 does not hold for proper reductions, but all the exceptions can be conveniently described. Let us do that now.

Let $C_1$ and $C_2$ be two vertex-disjoint cycles of length $n \geq 3$ with vertex-sets $\{u_1, u_2, \ldots, u_n\}$ and $\{v_1, v_2, \ldots, v_n\}$ (in order), respectively, and let $G_1$ be the graph obtained from the union of $C_1$ and $C_2$ by adding an edge joining $u_i$ and $v_i$ for each $i = 1, 2, \ldots, n$. We say that $G_1$ is a *planar ladder*. Let $G_2$ be the graph consisting of a cycle $C$ with vertex-set $\{u_1, u_2, \ldots, u_{2n}\}$ (in order), where $n \geq 2$ is an integer, and $n$ edges with ends $u_i$ and $u_{n+i}$ for $i = 1, 2, \ldots, n$. We say that $G_2$ is a *Möbius ladder*. A *ladder* is a planar ladder or a Möbius ladder. Let $G_1$ be a planar ladder as above on at least six vertices, and let $G_3$ be obtained from $G_1$ by deleting the edge $u_1u_2$ and contracting the edges $u_1v_1$ and $u_2v_2$. We say that $G_3$ is a *staircase*. Let $t \geq 2$ be an integer, and let $P$ be a path with vertices $v_1, v_2, \ldots, v_t$ in order. Let $G_4$ be obtained from $P$ by adding two distinct vertices $x, y$ and edges $xv_i$ and $yv_j$ for $i = 1, t$ and all even $i \in \{1, 2, \ldots, t\}$ and $j = 1, t$ and all odd $j \in \{1, 2, \ldots, t\}$.

Let $G_5$ be obtained from $G_4$ by adding the edge $xy$. We say that $G_5$ is an *upper prismoid*, and if $t \geq 4$, then we say that $G_4$ is a *lower prismoid*. A *prismoid* is a lower prismoid or an upper prismoid. We are now ready to state a strengthening of Theorem 6.3, proved in [43].

**Theorem 6.5.** *Let $H$, $G$ be bricks, where $H$ is isomorphic to a matching minor of $G$. Assume that $H$ is not isomorphic to $K_4$ or the prism, and $G$ is not a ladder, wheel, staircase or prismoid. Then a graph isomorphic to $H$ can be obtained from $G$ by repeatedly taking proper reductions in such a way that all the intermediate graphs are bricks not isomorphic to the Petersen graph.*

As a counterpart to Theorem 6.5, [43] describes the starting graphs for the generation process. Notice that $K_4$ is a wheel, a Möbius ladder, a staircase and an upper prismoid, and that the prism is a planar ladder, a staircase and a lower prismoid.

**Theorem 6.6.** *Let $G$ be a brick not isomorphic to $K_4$, the prism or the Petersen graph. Then $G$ has a matching minor isomorphic to one of the following seven graphs: the graph obtained from the prism by adding an edge, the lower prismoid on eight vertices, the staircase on eight vertices, the staircase on ten vertices, the planar ladder on ten vertices, the wheel on six vertices, and the Möbius ladder on eight vertices.*

If $H$ is a brick isomorphic to a matching minor of a brick $G$ and $G$ is a ladder, wheel, staircase or prismoid, then $H$ itself is a ladder, wheel, staircase or prismoid, and can be obtained from a graph isomorphic to $G$ by taking (improper) reductions in such a way that all intermediate graphs are bricks. Thus Theorems 6.5 and 6.6 imply Theorem 6.3. Theorems 6.5 and 6.6 were used to prove two results about minimal bricks [42], and to generate interesting examples of Pfaffian bricks. We will discuss some of those later.

McCuaig [34] proved an analogue of Theorem 6.5 for braces and used it in his proof of Theorem 4.2 in [35]. To state his result we need another exceptional class of graphs. Let $C$ be an even cycle with vertex-set $v_1, v_2, \ldots, v_{2t}$ in order, where $t \geq 2$ is an integer and let $G_6$ be obtained from $C$ by adding vertices $v_{2t+1}$ and $v_{2t+2}$ and edges joining $v_{2t+1}$ to the vertices of $C$ with odd indices and $v_{2t+2}$ to the vertices of $C$ with even indices. Let $G_7$ be obtained from $G_6$ by adding an edge $v_{2t+1}v_{2t+2}$. We say that $G_7$ is an *upper biwheel*, and if $t \geq 3$ we say that $G_6$ is a *lower biwheel*. A *biwheel* is a lower biwheel or an upper biwheel. McCuaig's result is as follows.

**Theorem 6.7.** *Let $H$, $G$ be braces, where $H$ is isomorphic to a matching minor of $G$. Assume that if $H$ is a planar ladder, then it is the largest planar ladder matching minor of $G$, and similarly for Möbius ladders, lower biwheels and upper biwheels. Then a graph isomorphic to $H$ can be obtained from $G$ by repeatedly taking proper reductions in such a way that all the intermediate graphs are braces.*

Actually, Theorem 6.7 follows from a stronger version of Theorem 6.5 proved in [43].

## 7. Crossing numbers and $k$-Pfaffian graphs

By a drawing $\Gamma$ of a graph in a surface $\Sigma$ we mean an immersion of $G$ in $\Sigma$ such that edges are represented by homeomorphic images of $[0, 1]$, not containing vertices in their interiors. Edges are permitted to intersect, but there are only finitely many intersections and each intersection is a crossing. For edges $e, f$ of a drawing $\Gamma$ let $cr(e, f)$ denote the number of times the edges $e$ and $f$ cross. For a perfect matching $M$ let $cr_\Gamma(M)$, or $cr(M)$ when $\Gamma$ is understood from the context, denote $\sum cr(e, f)$, where the sum is taken over all unordered pairs of distinct edges $e, f \in M$. The following theorem was proved by Norine [38]. The "if" part was known to Kasteleyn [23] and was proved by Tesler [56]. Norine's proof of that implication is different.

**Theorem 7.1.** *A graph $G$ is Pfaffian if and only if there exists a drawing of $G$ in the plane such that $cr(M)$ is even for every perfect matching $M$ of $G$.*

The theory of crossing numbers is fairly well-developed, but only few results involve parity of crossing numbers, and I am not aware of any about crossings of perfect matchings. The closest relative of Theorem 7.1 seems to be the following classical result of Hanani [20] and Tutte [62].

**Theorem 7.2.** *Let $\Gamma$ be a drawing of a graph in the plane such that $cr(e, f)$ is even for every two distinct non-adjacent edges $e, f$ of $G$. Then $G$ is planar.*

In fact, there is a deeper connection between the last two theorems. Norine [40] generalized Theorem 7.1 to a statement about the parity of self-intersections of different $T$-joins of a graph, and this generalization implies Theorem 7.2 as well as other results about crossing numbers. We omit the precise statement and instead refer the readers to [40].

A graph $G$ is *$k$-Pfaffian* if there exist orientations $D_1, D_2, \ldots, D_k$ of $G$ and real numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that $\sum_{i=1}^{k} \alpha_i \mathrm{sgn}_{D_i}(M) = 1$ for every perfect matching $M$ of $G$. Thus if $k$ is fixed and the orientations and coefficients as above are given, then the number of perfect matchings of $G$ can be calculated efficiently, using Lemma 2.1. The following was noted by Kasteleyn [23] and proved by Galluccio and Loebl [17] and Tesler [56].

**Theorem 7.3.** *Every graph that has an embedding in the orientable surface of genus $g$ is $4^g$-Pfaffian.*

In light of Theorem 7.1 one might speculate that 4-Pfaffian graphs are related to graphs drawn on the torus subject to the parity condition of Theorem 7.1. That is indeed true, as shown by Norine [41].

**Theorem 7.4.** *Every 3-Pfaffian graph is Pfaffian. A graph $G$ is 4-Pfaffian if and only if it can be drawn on the torus such that $cr(M)$ is even for every perfect matching $M$ of $G$.*

It is therefore sensible to conjecture a generalization to surfaces of arbitrary genus, as does Norine in [41]:

**Conjecture 7.5.** For a graph $G$ and integer $g \geq 0$ the following conditions are equivalent:

(1) There exists a drawing of $G$ on the orientable surface of genus $g$ such that $cr(M)$ is even for every perfect matching $M$ of $G$.

(2) The graph $G$ is $4^g$-Pfaffian.

(3) The graph $G$ is $(4^{g+1} - 1)$-Pfaffian.

Norine [41] has shown that every 5-Pfaffian graph is 4-Pfaffian, but his method breaks down after that.

## 8. Signs of edge-colorings

In this section we relate signs of edge-colorings (as in Penrose [46]) with "Pfaffian labelings", a generalization of Pfaffian orientations, whereby edges are labeled by elements of an Abelian group with an element of order two.

A graph $G$ is called *$k$-list-edge-colorable* if for every set system $\{S_e : e \in E(G)\}$ such that $|S_e| = k$ there exists a proper edge coloring $c$ with $c(e) \in S_e$ for every $e \in E(G)$. The following famous list-edge-coloring conjecture was suggested independently by various researchers and first appeared in print in [5].

**Conjecture 8.1.** Every $k$-edge-colorable graph is $k$-list-edge-colorable.

In a $k$-regular graph $G$ one can define an equivalence relation on $k$-edge colorings as follows. Let $c_1, c_2 \colon E(G) \to \{1, \ldots, k\}$ be two (proper) $k$-edge colorings of $G$. For $v \in V(G)$ let $\pi_v \colon \{1, \ldots, k\} \to \{1, \ldots, k\}$ be the permutation such that $\pi_v(c_1(e)) = c_2(e)$ for every $e \in E(G)$ incident with $v$, and let $c_1 \sim c_2$ if $\prod_{v \in V(G)} sgn(\pi_v) = 1$. Obviously $\sim$ is an equivalence relation on the set of $k$-edge colorings of $G$ and $\sim$ has at most two equivalence classes. We say that $c_1$ and $c_2$ *have the same sign* if $c_1 \sim c_2$ and we say that $c_1$ and $c_2$ *have opposite signs* otherwise.

A powerful algebraic technique developed by Alon and Tarsi [2] implies [1] that if in a $k$-edge-colorable $k$-regular graph $G$ all $k$-edge colorings have the same sign then $G$ is $k$-list-edge-colorable. In [14] Ellingham and Goddyn prove the following theorem.

**Theorem 8.2.** *In a $k$-regular planar graph all $k$-edge colorings have the same sign. Therefore every $k$-edge-colorable $k$-regular planar graph is $k$-list-edge-colorable.*

Goddyn [18] conjectured that Theorem 8.2 generalizes to Pfaffian graphs. This turned out to be true, even for the somewhat larger class of graphs that admit Pfaffian labelings. Let us introduce those graphs now.

Let $\Gamma$ be an Abelian multiplicative group, let 1 be the identity of $\Gamma$ and let $-1$ be some element of order two in $\Gamma$. Let $G$ be a graph with $V(G) = \{1, 2, \ldots, n\}$,

and let $D$ be the orientation of $G$ in which every edge $ij$ is oriented from $i$ to $j$, where $i < j$. Let us recall that $\mathrm{sgn}_D(M)$ was defined in Section 2. We say that $l\colon E(G) \to \Gamma$ is a *Pfaffian labeling* of $G$ if for every perfect matching $M$ of $G$, $\mathrm{sgn}_D(M) = \prod_{e\in M} l(e)$. We say that $G$ admits a *Pfaffian $\Gamma$-labeling* if there exists a Pfaffian labeling $l\colon E(G) \to \Gamma$ of $G$. We say that $G$ admits a *Pfaffian labeling* if $G$ admits a Pfaffian $\Gamma$-labeling for some Abelian group $\Gamma$ as above. It is easy to see that a graph $G$ admits a Pfaffian $\mathbb{Z}_2$-labeling if and only if $G$ admits a Pfaffian orientation. Note also that the existence of a Pfaffian labeling of a graph does not depend on the ordering of its vertices. The results of the remainder of this section are from [44].

**Theorem 8.3.** *Let $G$ be a $k$-regular graph with $V(G) = \{1, \ldots, 2n\}$. If $G$ admits a Pfaffian labeling then all $k$-edge-colorings of $G$ have the same sign.*

Using the theory of Alon and Tarsi mentioned above this implies a proof of Goddyn's conjecture:

**Corollary 8.4.** *Every $k$-edge-colorable $k$-regular graph that admits a Pfaffian labeling is $k$-list-edge-colorable.*

Theorem 8.3 has the following partial converse.

**Theorem 8.5.** *If a graph $G$ does not admit a Pfaffian labeling then there exist an integer $k$, a $k$-regular multigraph $G'$ whose underlying simple graph is a spanning subgraph of $G$ and two $k$-edge colorings of $G'$ of different signs.*

The above two theorems suggest the study of graphs that admit a Pfaffian labeling. First, there is an analogue of Theorem 6.2.

**Lemma 8.6.** *Let $\Gamma$ be an Abelian group. A matching covered graph $G$ admits a Pfaffian $\Gamma$-labeling if and only if each of its bricks and braces admits a Pfaffian $\Gamma$-labeling.*

Thus it suffices to characterize bricks and braces that have a Pfaffian labeling. The Petersen graph is not Pfaffian, but it admits a Pfaffian $\mu_4$-labeling, where $\mu_n$ is the multiplicative group of the $n^{\text{th}}$ roots of unity. Figure 2 shows an example of such a labeling. Using Theorems 6.3 and 6.7 it is not hard to show that the Petersen graph is the only brick or brace with that property. Using Lemma 8.6 we obtain:

**Theorem 8.7.** *A graph $G$ admits a Pfaffian labeling if and only if every brick and brace in its decomposition is either Pfaffian or isomorphic to the Petersen graph. If $G$ admits a Pfaffian $\Gamma$-labeling for some Abelian group $\Gamma$ then $G$ admits a Pfaffian $\mu_4$-labeling.*

The last result of this section characterizes graphs that admit a Pfaffian labeling in terms of their drawing in the projective plane. We say that a region $C$ of the projective plane is *a crosscap* if its boundary is a simple closed curve and its complement is a disc. We say that a drawing $\Gamma$ of a graph $G$ in the projective plane is *proper with*

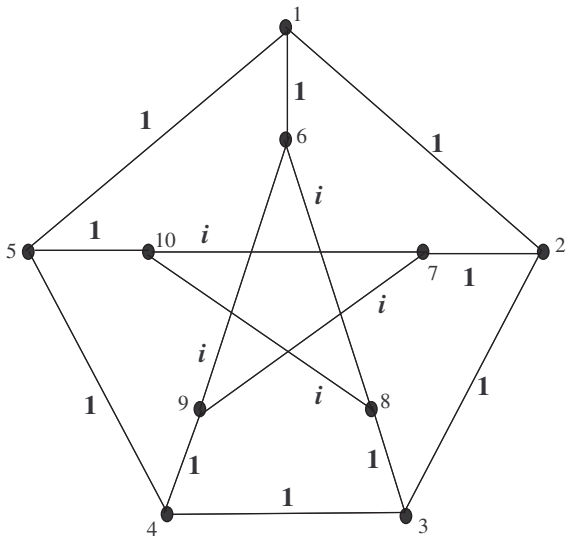



Figure 2. A $\mu_4$-labeling of the Petersen graph.

*respect to the crosscap C* if no vertex of $G$ is mapped to $C$ and for every edge $e \in E(G)$ intersecting $C$ and every crosscap $C' \subseteq C$ the image of $e$ intersects $C'$.

**Theorem 8.8.** *A graph $G$ admits a Pfaffian labeling if and only if there exists a crosscap $C$ in the projective plane and a proper drawing $\Gamma$ of $G$ in the projective plane with respect to $C$ such that $|M \cap S|$ and $cr_\Gamma(M)$ are even for every perfect matching $M$ of $G$, where $S \subseteq E(G)$ denotes the set of edges whose images intersect $C$.*

## 9. On characterizing Pfaffian graphs

Norine's theorem, Theorem 7.1, is a beautiful result, but, unfortunately, does not seem to help testing whether a graph is Pfaffian. Theorems 4.1 and 4.2 suggest two possible ways of characterizing Pfaffian graphs, but neither has been carried out, and there appear to be serious difficulties.

Fischer and Little [15] extended Theorem 4.1 as follows. A matching covered graph is *near bipartite* if it has two edges whose deletion makes the graph bipartite and matching covered. Let *Cubeplex* be the graph obtained from the (skeleton of the 3-dimensional) cube by subdividing three edges of a perfect matching and adding a vertex of degree three adjacent to the three resulting vertices, and let *Twinplex* be obtained from the Petersen graph by subdividing two edges that form an induced matching and joining the resulting vertices by an edge. This defines both graphs uniquely up to isomorphism. We say that a graph is a *weak matching minor* of another

if the first can be obtained from a matching minor of the second by contracting odd cycles and deleting all resulting loops and parallel edges.

**Theorem 9.1.** *A near bipartite graph is Pfaffian if an only if it has no weak matching minor isomorphic to $K_{3,3}$, Cubeplex or Twinplex.*

Let us say that a graph $G$ is *minimally non-Pfaffian* if $G$ is not Pfaffian but every proper weak matching minor of $G$ is. Thus $K_{3,3}$, Cubeplex and Twinplex are minimally non-Pfaffian, and so is the Petersen graph, as is easily seen. Fischer and Little (private communication) conjectured that those are the only minimally non-Pfaffian graphs; in other words they conjectured that upon adding the Petersen graph to the list of excluded weak matching minors, Theorem 9.1 holds for all graphs. Unfortunately, that is not the case. Here is an infinite family of minimally non-Pfaffian graphs [45].

Let $k \geq 2$, let $C_{2k+1}$ be the cycle of length $2k+1$ with vertices $1, 2, \ldots, 2k+1$ in order, and let $M$ be a matching in $C$, possibly empty. Let the graph $G(k, M)$ be defined by saying that its vertex-set is $\{u_1, u_2, \ldots, u_{2k+1}, v_1, v_2, \ldots, v_{2k+1}, w_1, w_2\}$ and that $G(k, M)$ has the following edges, where index arithmetic is taken modulo $2k+1$:

- $u_i v_i$ for all $i = 1, 2, \ldots, 2k+1$,
- $u_i u_{i+1}$ and $v_i w_2$ if $\{i, i+1\} \notin M$,
- $v_i w_1$ if $\{i-1, i\} \notin M$,
- $u_i v_{i+1}$ and $v_i u_{i+1}$ if $\{i, i+1\} \in M$.

Notice that the graph $G(2, \{\{1, 2\}, \{3, 4\}\})$ is isomorphic to Cubeplex.

**Theorem 9.2.** *For every integer $k \geq 2$ and every matching $M$ of $C_{2k+1}$ the graph $G(k, M)$ is minimally non-Pfaffian.*

Thus an extension of Theorem 9.1 to all graphs would have to involve infinitely many excluded weak matching minors. On the other hand, as noted in [45], the family $G(k, M)$ suggests a possible weakening of the weak matching minor ordering, and it is possible that if weak matching minor is replaced by this weakening, then the list of excluded might be finite (and of a reasonable size).

There is also the possibility of extending Theorem 4.2 to all graphs. That could be potentially very profitable, because it might lead to a polynomial-time recognition algorithm, but the prospects for that are not very bright. The class of planar graphs can be enlarged to a bigger class of Pfaffian graphs defined by means of surface embeddings. Let us say that an embedding of a graph $G$ in the Klein bottle is *cross-cap-odd* if every cycle $C$ in $G$ that does not separate the surface is odd if and only if it is 1-sided. If $G$ is embedded in the Klein bottle with all faces even (that is, bounded by a walk of even length), then the embedding is cross-cap-odd if and only if the 1-sided cycles are precisely the odd cycles in $G$. Please note that every planar graph can be embedded in the Klein bottle so that the embedding will be cross-cap-odd, and every embedding of a non-bipartite graph in the projective plane with all faces even may be regarded as a cross-cap-odd embedding in the Klein bottle. The following result,

proved in [39], resulted from earlier conversations of the author with Neil Robertson and Paul Seymour. By the remark above it implies Theorem 1.1.

**Theorem 9.3.** *Every graph that admits a cross-cap-odd embedding in the Klein bottle is Pfaffian.*

It may seem reasonable to expect an analogue of Theorem 4.2, something along the lines that every Pfaffian brick can be obtained from graphs that admit a cross-cap-odd embedding in the Klein bottle and a few sporadic exceptional graphs by means of certain composition operations. Unfortunately, the following construction of Norine seems to give a counterexample.

**Theorem 9.4.** *For every integer $n \geq 1$ there exists a Pfaffian brick that has a subgraph isomorphic to $K_n$.*

There is a chance that a notion analogous to tree-width can help us get around Norine's construction. A tree is *ternary* if all its vertices have degree one or three; the vertices of degree one are called *leaves*. A *matching decomposition* of a graph $G$ is a pair $(T, \tau)$, where $T$ is a ternary tree and $\tau$ is a bijection from the set of leaves of $T$ to $V(G)$. For an edge $e \in E(T)$ fix one of the two components of $T \setminus e$, and let $V_e$ be the set of all leaves of $T$ that belong to that component. We define the *width* of $e$ as the maximum, over all perfect matchings $M$ of $G$, of $|\delta(\tau(V_e)) \cap M|$. We define the *width* of $(T, \tau)$ as the maximum width of an edge of $T$, and we define the *matching-width* of a graph $G$ as the minimum width of a matching decomposition of $G$. The graphs Norine constructed in the proof of Theorem 9.4 all have low matching-width, and so that leaves open the possibility that Pfaffian graphs of high matching-width might exhibit more structure. Further, Norine [39] describes a polynomial-time algorithm to test whether an input graph $G$ is Pfaffian, assuming a matching decomposition of $G$ of bounded width is given as part of the input instance. Thus there is some hope, but at the moment it is not clear if these ideas can be turned into a meaningful theorem or a polynomial-time algorithm to test whether an input graph is Pfaffian. The following conjecture of Norine and the author [39], modeled after the excluded grid theorem of Robertson and Seymour [48] (see also [11], [49]), seems relevant.

**Conjecture 9.5.** There exists a function $f$ such that every graph of matching-width at least $f(k)$ has a matching minor isomorphic to the $2k \times 2k$ grid.

## References

[1] Alon, N., Restricted colorings of graphs. In *Surveys in Combinatorics*, London Math. Soc. Lecture Note Ser. 187, Cambridge University Press, Cambridge 1993, 1–33.

[2] Alon, N., and Tarsi, M., Colorings and orientations of graphs. *Combinatorica* **12** (1992), 125–134.

[3] Birkhoff, G., Three observations on linear algebra. *Univ. Nac. Tucumán. Revista A.* **5** (1946), 147–151 (in Spanish).

[4] Bollobás, B., *Modern graph theory*. Grad. Texts in Math. 184, Springer-Verlag, New York 1998.

[5] Bollobás, B., and Harris, A. J., List-colorings of graphs. *Graphs Combin.* **1** (1985), 115–127.

[6] Brualdi, R. A., and Shader, B. L., *Matrices of sign-solvable linear systems*. Cambridge Tracts in Math. 116, Cambridge University Press, Cambridge 1995.

[7] de Carvalho, M. H., Lucchesi, C. L., and Murty, U. S. R., On a conjecture of Lovász concerning bricks. II. Bricks of finite characteristic. *J. Combin. Theory B* **85** (2002), 137–180.

[8] de Carvalho, M. H., Lucchesi, C. L., and Murty, U. S. R., On the number of dissimilar Pfaffian orientations of graphs. *Theor. Inform. Appl.* **39** (2005), 93–113.

[9] de Carvalho, M. H., Lucchesi, C. L., and Murty, U. S. R., How to build a brick. *Discrete Math.*, to appear

[10] Diestel, R., *Graph Theory*. Grad. Texts in Math. 173, Springer-Verlag, Berlin 2005.

[11] Diestel, R., Jensen, T. R., Gorbunov, K. Yu., and Thomassen, C., Highly connected sets and the excluded grid theorem. *J. Combin. Theory Ser. B* **75** (1999), 61–73.

[12] Edmonds, J., Maximum matching and a polyhedron with (0, 1) vertices. *J. Res. Nat. Standards B* **69** (1965), 125–130.

[13] Edmonds, J., Lovász, L., and Pulleyblank, W. R., Brick decompositions and matching rank of graphs. *Combinatorica* **2** (1982), 247–274.

[14] Ellingham, M., and Goddyn, L., List edge colourings of some 1-factorable multigraphs. *Combinatorica* **16** (1996), 343–352.

[15] Fischer, I., and Little, C., A characterisation of Pfaffian near bipartite graphs. *J. Combin. Theory Ser. B* **82** (2001), 175–222.

[16] Fisher, M. E., Statistical mechanics of dimers on a plane lattice. *Phys. Rev.* **124**, 1664–1672.

[17] Galluccio, A., and Loebl, M., On the theory of Pfaffian orientations I. Perfect matchings and permanents. *Electron. J. Combin.* **6** (1999).

[18] Goddyn, L., private communication.

[19] Grötschel, M., Lovász, L., and Schrijver, A., *Geometric algorithms and combinatorial optimization.* Algorithms Combin. 2, Springer-Verlag, Berlin 1993.

[20] Hanani, H. (Chaim Chojnacki), Über wesentlich unplättbare Kurven im dreidimensionalen Raume. *Fund. Math.* **23** (1934), 135–142.

[21] Kasteleyn, P. W., The statistics of dimers on a lattice. I. The number of dimer arrangements on a quadratic lattice. *Physica* **27** (1961), 1209–1225.

[22] Kasteleyn, P. W., Dimer statistics and phase transitions. *J. Math. Phys.* **4** (1963), 287–293.

[23] Kasteleyn, P. W., Graph theory and crystal physics. In *Graph Theory and Theoretical Physics* (ed. by F. Harary), Academic Press, New York 1967, 43–110.

[24] Klee, V., Ladner, R., and Manber, R., Sign-solvability revisited. *Linear Algebra Appl.* **59** (1984), 131–158.

[25] Lieb, E. H., and Loss, M., Fluxes, Laplacians, and Kasteleyn's theorem. *Duke Math. J.* **71** (1993), 337–363.

[26] Little, C. H. C., Kasteleyn's theorem and arbitrary graphs. *Canad. J. Math.* **25** (1973), 758–764.

[27] Little, C. H. C., A characterization of convertible (0, 1)-matrices. *J. Combin. Theory Ser. B* **18** (1975), 187–208.

[28] Lovász, L., On chromatic number of finite set-systems. *Acta Math. Acad. Sci. Hungar.* **19** (1968), 59–67.

[29] Lovász, L., Ear-decompositions of matching covered graphs. *Combinatorica* **2** (1983), 395–407.

[30] Lovász, L., Some algorithmic problems on lattices. In *Theory of Algorithms* (ed. by L. Lovász and E. Szemerédi), Colloq. Math. Soc. Janos Bolyai 44, János Bolyai Mathematical Society, Budapest 1985, 323–337.

[31] Lovász, L., Matching structure and the matching lattice. *J. Combin. Theory Ser. B* **43** (1987), 187–222.

[32] Lovász, L., and Plummer, M., *Matching theory*. North-Holland Math. Stud. 121, Ann. of Discrete Math. 29, North-Holland, Amsterdam 1986.

[33] McCuaig, W., Even dicycles. *J. Graph Theory* **35** (2000), 46–68.

[34] McCuaig, W., Brace generation. *J. Graph Theory* **38** (2001), 124–169.

[35] McCuaig, W., Pólya's permanent problem. *Electron. J. Combin.* **11** (2004), 83pp.

[36] McCuaig, W., Robertson, N., Seymour, P. D., and Thomas, R., Permanents, Pfaffian orientations, and even directed circuits (Extended abstract). In *Proceedings of the 29th annual Symposium on the Theory of Computing* (*STOC*), ACM Press, New York 1997, 402–405.

[37] Muir, T., *The theory of determinants*. MacMillan and Co., London 1906.

[38] Norine, S., Drawing Pfaffian graphs. Submitted.

[39] Norine, S., Matching structure and Pfaffian orientations of graphs. Ph.D. dissertation, Georgia Institute of Technology, 2005.

[40] Norine, S., Pfaffian graphs, $T$-joins and crossing numbers. Submitted.

[41] Norine, S., Drawing 4-Pfaffian graphs on the torus. Submitted.

[42] Norine, S., and Thomas, R., Minimal bricks. *J. Combin. Theory Ser. B* **96** (4) (2006), 505–513.

[43] Norine, S., and Thomas, R., Generating bricks. Submitted.

[44] Norine, S., and Thomas, R., Pfaffian labelings and signs of edge-colorings. Submitted.

[45] Norine, S., and Thomas, R., On minimally non-Pfaffian graphs. Manuscript.

[46] Penrose, R., Applications of negative-dimensional tensors. In *Combinatorial Mathematics and its Applications* (ed. by D. J. A. Welsh), Academic Press, London 1971, 221–244.

[47] Pólya, G., Aufgabe 424. *Arch. Math. Phys. Ser.* **20** (1913), 271.

[48] Robertson, N., and Seymour, P. D., Graph Minors V. Excluding a planar graph. *J. Combin. Theory Ser. B* **41** (1986), 92–114.

[49] Robertson, N., Seymour, P. D., and Thomas, R., Quickly excluding a planar graph. *J. Combin. Theory Ser. B* **62** (1994), 323–348.

[50] Robertson, N., Seymour, P. D., and Thomas, R., Permanents, Pfaffian orientations, and even directed circuits. *Math. Ann.* **150** (1999), 929–975.

[51] Samuelson, P. A., *Foundations of economic analysis*. Harvard University Press, Cambridge, MA, 1947.

[52] Seymour, P. D., On the two-colouring of hypergraphs. *Quart. J. Math. Oxford* **25** (1974), 303–312.

[53] Seymour, P. D., and Thomassen, C., Characterization of even directed graphs. *J. Combin. Theory Ser. B* **42** (1987), 36-45.

[54] Tait, P. G., Note on a theorem in geometry of position, *Trans. Roy. Soc. Edinburgh* **29** (1880), 657–660.

[55] Temperley, H. N. V., and Fisher, M. E., Dimer problem in statistical mechanics—an exact result. *Phil. Mag.* **6** (1961), 1061–1063.

[56] Tesler, G., Matching in graphs on non-orientable surfaces. *J. Combin. Theory Ser. B* **78** (2000), 198–231.

[57] Thomassen, C., Even cycles in directed graphs. *European J. Combin.* **6** (1985), 85–89.

[58] Thomassen, C., Sign-nonsingular matrices and even cycles in directed graphs. *Linear Algebra Appl.* **75** (1986), 27–41.

[59] Thomassen, C., The even cycle problem for directed graphs. *J. Amer. Math. Soc.* **5** (1992), 217–229.

[60] Thomassen, C., The even cycle problem for planar digraphs. *J. Algorithms* **15** (1993), 61–75.

[61] Tutte, W. T., The factorization of linear graphs. *J. London Math. Soc.* **22** (1947), 107–111.

[62] Tutte, W. T., Toward a theory of crossing numbers, *J. Combin. Theory Ser. B* **8** (1971), 45–53.

[63] Valiant, L. G., The complexity of computing the permanent. *Theoret. Comput. Sci.* **8** (1979), 189–201.

[64] Vazirani, V. V., and Yannakakis, M., Pfaffian orientations, 0-1 permanents, and even cycles in directed graphs. *Discrete Appl. Math.* **25** (1989), 179–190.

[65] West, D., *Introduction to Graph Theory*. Prentice-Hall, Upper Saddle River, NJ, 2001.

School of Mathematics, Georgia Tech, Atlanta, GA 30332-0160, U.S.A.
E-mail: thomas@math.gatech.edu

# Determinant versus permanent

Manindra Agrawal

**Abstract.** We study the problem of expressing permanents of matrices as determinants of (possibly larger) matrices. This problem has close connections to the complexity of arithmetic computations: the complexities of computing permanent and determinant roughly correspond to arithmetic versions of the classes NP and P respectively. We survey known results about their relative complexity and describe two recently developed approaches that might lead to a proof of the conjecture that the permanent can only be expressed as the determinant of exponential-sized matrices.

**Mathematics Subject Classification (2000).** Primary 68Q17; Secondary 68W30.

**Keywords.** Arithmetic computation, complexity classes, determinant, permanent.

## 1. Introduction

The determinant of square matrices plays a fundamental role in linear algebra. It is a linear function of rows (and columns) of the matrix, and has several nice interpretations. Geometrically it is the volume of the parallelopied defined by rows (or columns) of the matrix, and algebraically it is the product of all eigenvalues, with multiplicity, of the matrix. It also satisfies a number of other interesting properties, e.g., it is multiplicative, invariant under linear combinations of rows (and columns) etc. The *permanent* of a square matrix is a number that is defined in a way similar to the determinant. For a matrix $X = [x_{i,j}]_{1\le i,j\le n}$,

$$\operatorname{per} X = \sum_{\pi\in S_n}\prod_{i=1}^{n} x_{i,\pi(i)},$$

where $S_n$ is the symmetric group on $n$ elements. The only difference to the determinant is in the signs of terms:

$$\det X = \sum_{\pi\in S_n}\operatorname{sgn}(\pi)\cdot\prod_{i=1}^{n} x_{i,\pi(i)},$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

where $\mathrm{sgn}(\pi) \in \{1, -1\}$ is the sign of the permutation $\pi$.[1] Despite the similarity in definition, the permanent has much fewer properties than the determinant. No nice geometric or algebraic interpretation is known for permanent; and it is neither multiplicative nor invariant under linear combinations of rows or columns. Perhaps for this reason, permanents did not get much attention until the late 1970s, and just about everything known about it until then is in the book [10]. In 1979, Leslie Valiant [24] completely changed the view on permanents by showing that the complexity of computing permanent precisely captures the arithmetic version of the class NP, called VNP. Since then, properties of the permanent have been extensively studied by complexity theorists.

One of the most natural questions about permanents is to investigate its relationship with determinants. It is easy to see that the determinant of a matrix $X$ can be expressed as the permanent of a related matrix $\hat{X}$ whose entries are 0, 1, or $x_{i,j}$s and which is of size $O(n)$ (set up entries of $\hat{X}$ such that $\det \hat{X} = \det X$ and the product corresponding to every permutation that has an even cycle is zero). For the converse, the best known bound on the size of a matrix $\hat{X}$ whose entries are constants and $x_{i,j}$s, and for which $\det \hat{X} = \mathrm{per}\, X$ is $2^{\Omega(n)}$. This suggests that the complexity of computing the permanent is much higher than that of the determinant. Although widely believed, this remains a conjecture. This conjecture has a close connection to the conjectured separation of arithmetic NP from arithmetic P (the class of all functions that can be efficiently computed by arithmetic operations, see next section for a precise definition). It is known that the complexity of determinant is close to the complexity of arithmetic P: every function computed by $n$ arithmetic operations can be expressed as determinant of a matrix of size $n^{O(\log n)}$. This lends more importance to the problem of settling the conjecture.

There have been some attempts to answer the conjecture positively [14], [6], [15] [8]. A sequence of arithmetic operations can be modeled as an *arithmetic circuit*, and the size of an arithmetic circuit is the number of arithmetic operations in the sequence. In [8], *monotone* circuits were considered, these are circuits in which no constant is negative. For computability by such restricted circuits, an exponential lower bound was shown for the complexity of permanent. A different restriction on arithmetic circuits is that of depth – the number of layers of operations. These circuits were considered in [14], [19], [6] and lower bounds were shown for the complexity of computing permanent by depth three circuits. Finally, [15] considers yet another restriction. In this restriction, every gate of the circuit is required to compute a multilinear polynomial. A superpolynomial lower bound is shown on *formulas* (circuits with outdegree one) of this kind computing permanent.

All the above lower bounds hold for very restricted settings, and the techniques used do not seem to generalize. Over the last few years, however, two new tech-

[1] Both permanent and determinant are special forms of *immanents* defined as $\mathrm{imm}_\chi X = \sum_{\pi \in S_n} \chi(\pi) \cdot \prod_{i=1}^{n} x_{i,\pi(i)}$ where $\chi : S_n \mapsto \mathbb{C}$ is a character of $S_n$. For the permanent, $\chi = \mathrm{id}$ and for the determinant, $\chi$ equals the sign of the permutation.

niques have been proposed that hold some promise. The first of these was proposed by Mulmuley and Sohoni [11]. They transform the problem to algebraic geometry domain where it is reduced to showing that the permanent polynomial does not lie in the closure of a certain *orbit* of the determinant polynomial.

The second approach was proposed by Kabanets and Impagliazzo [9]. They reduced the problem to that of finding a deterministic, subexponential-time algorithm for the *Identity Testing*. The Identity Testing problem is defined as follows: given an arithmetic circuit computing polynomial $p$ in $n$ variables, test if $p = 0$. There exist several randomized polynomial-time algorithms for solving this. Kabanets and Impagliazzo show that *any* deterministic, subexponential-time algorithm for the problem will imply either a superpolynomial lower bound for arithmetic circuits computing permanent, or a boolean lower bound on the class NEXP. This connection was strengthened in [1] to show that if there exist special kinds of deterministic algorithms for testing identities given by superconstant depth arithmetic circuits, then permanent requires superpolynomial sized arithmetic circuits.

In this article, we will describe the known results on lower bounds on permanent as well as the two new approaches outlined above.

## 2. Definitions

$\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are respectively the fields of rational, real, and complex numbers.

An arithmetic circuit over a field $F$ is a directed, acyclic graph with labelled vertices. Vertices of indegree zero are labelled with either a variable $x_i$ or a constant from $F$. Vertices labelled with variables are called *input gates*. The remaining vertices are labelled with either '+' or '$*$' and are called *addition* or *multiplication gates* respectively. Vertices with outdegree zero are also called *output gates*. We restrict our attention to circuits with exactly one output gate. The *fanin* of a gate equals the number of edges incident to the gate. In this article, gates have unbounded fanin when not mentioned otherwise. The *size* of a circuit $C$ equals the number of gates in it. The *depth* of a circuit $C$ equals the length of the longest path from an input gate to an output gate. The *degree* of $C$ is inductively defined as follows: the degree of an input gate is one or zero depending on whether it is labelled by a variable or constant; the degree of an addition gate is the maximum of the degree of the gates whose edges are incident to the gate; the degree of a multiplication gate is the sum of the degrees of the gates whose edges are incident to the gate; finally, the degree of $C$ is the degree of its output gate.

An arithmetic circuit $C$ computes a polynomial as follows. The polynomial computed at an input gate equals the label of the gate. For any other gate $g$, let $g_1, \ldots, g_k$ be all the gates that have an edge incident to $g$ and let $p_{g_i}$ be the polynomial computed at gate $g_i$. Then the polynomial computed at the gate $g$ equals $\sum_{i=1}^{k} p_{g_i}$ if $g$ is an addition gate, and equals $\prod_{i=1}^{k} p_{g_i}$ if $g$ is a multiplication gate. The polynomial

computed by the circuit is the polynomial computed at its output gate.

Let $\{p_n\}_{n>0}$ be a family of polynomials with $p_n$ a polynomial of degree $d(n)$ in $n$ variables. A circuit family $\{C_n\}_{n>0}$ is said to compute $\{p_n\}$ if for every $n$, the polynomial computed by $C_n$ equals $p_n$. In the following we shall simply write $\{p_n\}$ for the family $\{p_n\}_{n>0}$.

*Arithmetic branching programs* are a restricted form of arithmetic circuits in which every multiplication gate has fanin exactly two, and in addition at least one of the two gates, from which edges are incident to the multiplication gate, is an input gate. Such circuits are also called *skew* circuits.

The class $\mathrm{VP}_F$, the arithmetic analog of class P, is defined to be the set of polynomial families $\{p_n\}$ over a field $F$ such that (1) each $p_n$ is of degree $n^{O(1)}$, and (2) there exists a circuit family $\{C_n\}$ computing $\{p_n\}$ such that $C_n$ is of size $n^{O(1)}$.[2] The class $\mathrm{VNP}_F$, the arithmetic analog of class NP, is defined to be the set of polynomial families $\{p_n\}$ over a field $F$ such that (1) each $p_n$ is of degree $n^{O(1)}$, and (2) there exists a family of polynomials $\{q_n\} \in \mathrm{VP}_F$ such that for every $n$,

$$p_n(x_1, x_2, \ldots, x_n) = \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \cdots \sum_{y_m=0}^{1} q_{n+m}(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$$

with $m = n^{O(1)}$.[3]

Given two polynomials $p(x_1, x_2, \ldots, x_n)$ and $q(y_1, y_2, \cdots, y_m)$ over a field $F$, we say that $p$ is a *projection* of $q$ if $p(x_1, x_2, \ldots, x_n) = q(\alpha_1, \alpha_2, \cdots, \alpha_m)$ where each $\alpha_i \in F \cup \{x_1, x_2, \ldots, x_n\}$. Given two polynomial families $\{p_n\}$ and $\{q_n\}$, we say that $\{p_n\}$ is a *p-projection* of $\{q_n\}$ if for every $n$ there exists an $m = n^{O(1)}$ such that $p_n$ is a projection of $q_m$.[4]

Let $\mathrm{per}_F = \{\mathrm{per}_{F,n}\}$ and $\mathrm{det}_F = \{\mathrm{det}_{F,n}\}$ denote the families of permanent and determinant polynomials over a field $F$ respectively. Both these families contain polynomials in $n^2$ variables for each $n$.

Valiant [24] proved that:

**Theorem 2.1** ([24]). *For any $F$, $\mathrm{per}_F \in \mathrm{VNP}_F$. In addition, for any $F$, $\mathrm{char}(F) \neq 2$, for any polynomial family $\{p_n\}$ in $\mathrm{VNP}_F$, $\{p_n\}$ is a p-projection of $\mathrm{per}_F$.*

So permanent is as hard to compute as any polynomial family in VNP. In contrast, determinant can be computed efficiently. A nice characterization of determinant was shown in [4], [21], [25]:

**Theorem 2.2** ([4], [21], [25]). *For any $F$, $\mathrm{det}_F$ can be computed by polynomial-sized arithmetic branching programs. In addition, for any $F$ and for any polynomial*

[2] In addition, circuit $C_n$ must be efficiently computable given $1^n$. This property does not seem to play any role in obtaining lower bounds.

[3] The class $\#\mathrm{P}_F$ is the 'functional' version of the class $\mathrm{VNP}_F$: a polynomial family $\{p_n\} \in \mathrm{VNP}_F$ belongs to $\#\mathrm{P}_F$ when for each $n$, $p_n$ is viewed as a map from $F^n$ to $F$.

[4] Again, given $1^n$, the projection specified by $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ should be efficiently computable.

*family* $\{p_n\}$ *computed by polynomial-sized arithmetic branching programs,* $\{p_n\}$ *is a p-projection of* $\det_F$.

In fact, all families in VP are *almost* p-projections of determinant.

**Theorem 2.3** ([23]). *Let C be a circuit of size s computing a polynomial of degree d. There exists another circuit computing the same polynomial of size* $s^{O(1)}$ *and depth* $O(\log s + \log d)$.

**Corollary 2.4.** *Any circuit family in* $\mathrm{VP}_F$ *can be computed by circuit families of polynomial size and logarithmic depth.*

**Corollary 2.5.** *For every circuit family* $\{p_n\} \in \mathrm{VP}_F$ *and for every n,* $p_n$ *is a projection of* $\det_{F,m}$ *where* $m = n^{O(\log n)}$.

The above characterizations of complexities of determinant and permanent imply that, in order to separate $\mathrm{VP}_F$ from $\mathrm{VNP}_F$, it is enough to show that $\mathrm{per}_F$ is not an almost p-projection of $\det_F$ (in the sense above).

## 3. Known lower bounds on permanent

Lower bounds are known on permanent only for restricted circuits. In this section, we describe important lower bounds of this kind. Three major restrictions have been considered for proving such lower bounds: *monotone* circuits, *constant depth* circuits, and *multilinear* formulas.

**3.1. Monotone circuits.** A circuit over $\mathbb{Q}$ or $\mathbb{R}$ is *monotone* if all the constants in the circuit are non-negative. This is a very restricted class of circuits since *no* cancellations can occur in it. Jerrum and Snir [8] showed that any monotone circuit family that computes permanent must have exponential size.

**3.2. Constant depth circuits.** In this restriction, the depth of a circuit family is fixed, i.e., it is independent of $n$. Permanent (or any polynomial of degree $n^{O(1)}$ for that matter) can be computed by an exponential size depth two circuit family. Conversely, it is easy to see that any depth two circuit family computing permanent must have exponential size.

Depth three circuit families are, however, non-trivial. A depth three circuit can be of the form "sum-of-products-of-sums" or "product-of-sums-of-products." The latter form can easily be seen to require exponential size to compute permanent (the topmost multiplication gate can be shown to be redundant transforming the circuit to a depth two circuit). Circuit families of the first form are powerful: Ben Or observed that they can efficiently compute all symmetric polynomials of degree $n^{O(1)}$ over fields of characteristic zero.

The best known lower bound in fields of characteristic zero is by Shpilka and Wigderson [19] who proved that permanent (and determinant) requires a circuit family of size $\Omega(n^2)$. Their idea is to consider the space spanned by all partial derivatives of the polynomials computed at each gate of a given circuit. They show that for a depth three circuit with small size, the space spanned by the derivatives of the output polynomial would be of small dimension while the space spanned by derivatives of permanent is of large dimension.

Over finite fields, the situation is better. Grigoriev and Razborov [6] showed an exponential lower bound on the size of depth three circuit families computing determinant and permanent. Their approach was to show that polynomial computed by a depth three circuit of small size can be 'approximated' by a low-degree polynomial (approximated in the sense that the two polynomials agree on a large set of points from the field). Then they observed that determinant and permanent cannot be approximated by low-degree polynomials.

**3.3. Multilinear formulas.** *Multilinear formulas* are circuits such that (1) the out-degree of every gate is at most one, and (2) the polynomial computed at every gate is multilinear. Such circuits have severely limited multiplication gates – the polynomials input to a multiplication gate must be over disjoint sets of variables. Using a combination of partial derivative technique and random restrictions (setting some input variables to random values), Raz [15] proved a lower bound of $n^{\Omega(\log n)}$ on the size of families of multilinear formulas computing permanent and determinant.

## 4. The algebraic geometry approach

Mulmuley and Sohoni [11] have offered a completely new approach to the problem of proving a lower bound on permanent for unrestricted circuits by transforming the problem to geometric settings. In this section, we give a brief overview of their approach.

Suppose, for $F = \mathbb{Q}$, $\mathrm{per}_{F,n}$ is a projection of $\det_{F,m}$, $m > n$. Define $\widehat{\mathrm{per}}_{F,m} = x_{m^2}^{m-n} \cdot \mathrm{per}_{F,n}$. It follows that $\widehat{\mathrm{per}}_{F,m}$ is also a projection of $\det_{F,m}$ (just multiply all constants of the projection by $x_{m^2}$). This can be written as

$$\widehat{\mathrm{per}}_{F,m}(x_1, x_2, \ldots, x_{m^2}) = A \cdot \det_{F,m} = \det_{F,m}((x_1, x_2, \ldots, x_{m^2}) \cdot A),$$

where $A$ is an $m^2 \times m^2$ matrix over $\mathbb{Q}$. The matrix $A$ is singular whenever $m > n$ since the variables $x_{n^2+1}, \ldots, x_{m^2-1}$ do not occur in $\widehat{\mathrm{per}}_{F,m}$. Let $A_{\bar{\varepsilon}}$ be a slight 'perturbation' of $A$ obtained by adding $\varepsilon_{i,j}$ to the $(i, j)$th entry of $A$. For nearly all values of $\bar{\varepsilon}$ close to zero, $A_{\bar{\varepsilon}}$ is non-singular and the polynomial $A_{\bar{\varepsilon}} \cdot \det_{F,m}$ approximates the polynomial $\widehat{\mathrm{per}}_{F,m}$ very well (all the coefficients of two polynomials are close to each other). Now consider the space $V = \mathbb{C}^M$ with $M = \binom{m^2+m-1}{m}$. Every homogeneous polynomial of degree $m$ in $m^2$ variables can be viewed as a point in this space (degree $m$ monomials

forming the basis). So both $\det_{F,m}$ and $\widehat{\text{per}}_{F,m}$ are points in $V$ (since $F = \mathbb{Q}$ and both polynomials are of degree $m$ in $m^2$ variables). Let $O$ be the orbit of $\det_{F,m}$ under the action of $\text{GL}_{m^2}(\mathbb{C})$, i.e.,

$$O = \{B \cdot \det_{F,m} \mid B \text{ is an invertible matrix over } \mathbb{C}\}.$$

Set $O$ can be viewed as a set of points in $V$. The above argument shows the following:

**Lemma 4.1** ([11]). *If* $\text{per}_{F,n}$ *is a projection of* $\det_{F,m}$ *then the point corresponding to* $\widehat{\text{per}}_{F,m}$ *in* $V$ *lies in the closure of the set* $O$ *in* $V$. *Conversely, if* $\widehat{\text{per}}_{F,m}$ *lies in the closure of* $O$ *then* $\text{per}_{F,n}$ *can be approximated by projections of* $\det_{F,m}$ *to any desired accuracy.*

This (near) characterization is the starting point of their approach. Instead of $V$, we can work in the projective space $P(V)$ too since both the polynomials are homogeneous. The same near characterization holds in $P(V)$ as well with $\text{GL}_{m^2}(\mathbb{C})$ replaced by $\text{SL}_{m^2}(\mathbb{C})$, the group of all matrices with determinant 1. The advantage of working in $P(V)$ is that the closure of $O$ (under the classical Euclidean topology) coincides with the closure of $O$ under *Zariski topology* [12]. In Zariski topology, there is the well-studied notion of *stability* that captures this problem: $\det_{F,m}$ is $\widehat{\text{per}}_{F,m}$*-stable under* $\text{SL}_{m^2}(\mathbb{C})$ if $\widehat{\text{per}}_{F,m}$ lies in the closure of the orbit $O$ (we abuse notation here by using the same names for polynomials and sets in $P(V)$ as for the corresponding ones in $V$).

Points in the orbit $O$ have a useful property. For any point $p \in P(V)$, let

$$G_p = \{A \in \text{SL}_{m^2}(\mathbb{C}) \mid A \cdot p = p\}.$$

The group $G_p$ is called the *stabilizer* of $p$.

**Lemma 4.2.** *For any point* $p \in O$, $G_p$ *is a conjugate of* $G_{\det_{F,m}}$.

*Proof.* Let $p = B \cdot \det_{F,m} \in O$. Then $G_p = B \cdot G_{\det_{F,m}} \cdot B^{-1}$. □

Suppose the orbit of the polynomial $\widehat{\text{per}}_{F,m}$ under $\text{SL}_{m^2}(\mathbb{C})$ is a closed set (such polynomials are called *stable*). Let $Q$ be the orbit of $\widehat{\text{per}}_{F,m}$ under $\text{SL}_{m^2}(\mathbb{C})$. By *Luna's slice theorem*, there is a neighborhood $N$ of $Q$ such that for any point $p \in N$, $G_p$ is a conjugate of a subgroup of $G_{\widehat{\text{per}}_{F,m}}$. Since the closure of $O$ contains $\widehat{\text{per}}_{F,m}$, there is a point in $N$, say $q$, such that $q = B \cdot \det_{F,m}$. This means $G_q$ is a conjugate of $G_{\det_{F,m}}$. Therefore, $G_{\det_{F,m}}$ is a conjugate of a subgroup of $G_{\widehat{\text{per}}_{F,m}}$. On the other hand, it is well known that $G_{\det_{F,m}}$ is 'larger' than $G_{\widehat{\text{per}}_{F,m}}$: $G_{\det_{F,m}}$ is characterized by the transformations of the kind $X \mapsto A \cdot X \cdot B^{-1}$ where $A, B \in \text{GL}_m(\mathbb{C})$ while $G_{\widehat{\text{per}}_{F,m}}$ is characterized by the transformations of the kind $X \mapsto A \cdot X \cdot B^{-1}$ where $A, B \in \text{GL}_m(\mathbb{C})$ and both $A$ and $B$ are either diagonal or permutation matrices. Therefore, $G_{\det_{F,m}}$ cannot be a conjugate of a subgroup of $G_{\widehat{\text{per}}_{F,m}}$. (This is a rough argument; to make it precise, more work is needed.)

Unfortunately, $\widehat{\text{per}}_{F,m}$ is *not* stable (interestingly, $\text{per}_{F,n}$ is stable in the smaller dimensional space defined by degree $n$ homogeneous polynomials in $n^2$ variables; the translation to higher dimensional space ruins the stability). Mulmuley and Sohoni define the notion of *partial stability* and show that $\widehat{\text{per}}_{F,m}$ is partially stable. Now their aim is to make the above argument work even for partially stable points. A more detailed explanation of their approach is in [16].

## 5. The derandomization approach

Kabanets and Impagliazzo [9] have discovered another new approach for proving lower bounds on permanent. Unlike the previous one, this approach is based on arithmetic circuits. In this section we outline their approach and its variation in [1].

The *Identity Testing* problem is defined as follows: given an arithmetic circuit $C$ over a field $F$ as input, decide if the polynomial computed by the circuit is the zero polynomial. This is a classical problem in computational algebra and there exist several randomized polynomial-time algorithms for it. Perhaps the simplest one is by Schwartz and Zippel [17], [26]: randomly choose values for variables of $C$ from a set in $F$ of size $2d$, here $d$ is the degree of $C$ (if $|F| < 2d$ then extend $F$ slightly); output ZERO if $C$ evaluates to zero, otherwise NON-ZERO. An easy argument shows that this test is correct with probability at least $\frac{1}{2}$ when $C$ computes a non-zero polynomial and always correct when $C$ computes a zero polynomial.

Kabanets and Impagliazzo show that if there exists a deterministic subexponential ($= 2^{n^{o(1)}}$) time algorithm for solving Identity Testing problem then at least one of the following two lower bounds hold:

1. NEXP requires superpolynomial sized boolean circuits.
2. Permanent requires superpolynomial sized arithmetic circuits.

To see this, suppose that permanent has polynomial sized arithmetic circuits for some field $F$ of characteristic different from two. Consider a non-deterministic machine that, on input $1^n$, guesses the circuit that computes $\text{per}_{F,n}$ and verifies it to be correct. It does this by inductively verifying that the circuit, under appropriate settings of its inputs, computes $\text{per}_{F,n-1}$ correctly and then verifying the equation for $\text{per}_{F,n}$ that expresses it in terms of $\text{per}_{F,n-1}$. Verifying the equation is an instance of Identity Testing problem and so can be done in subexponential time by assumption. Therefore, given any matrix $A \in F^{n^2}$, per $A$ can be computed in non-deterministic subexponential time. Now assume that NEXP has polynomial sized boolean circuits. By [3], [22], it follows that NEXP $\subseteq$ P$^{\#\text{P}}$. Since the complexity of #P is exactly the complexity of computing permanent, it follows that NEXP is in non-deterministic subexponential time contradicting the non-deterministic time hierarchy theorem [18].

This result falls short of pointing a way for proving lower bounds on permanent – besides finding a deterministic algorithm for Identity Testing, one needs to assume

NEXP has polynomial sized boolean circuits which is very unlikely to be true. However, it *does* point to a connection between Identity Testing problem and permanent lower bounds. This connection was strengthened in [1] by defining *pseudo-random generators* for arithmetic circuits. Pseudo-random generators in the boolean settings have been studied intensively (see, e.g., [5], [13], [7], [20]). It is known that constructing pseudo-random generators is equivalent to proving lower bounds in the boolean settings. In [1], pseudo-random generators are defined in arithmetic settings and a similar equivalence is observed.

Let $\mathcal{AC}_F$ be the class of all arithmetic circuits over $F$ and $\mathcal{A}_F \subseteq \mathcal{AC}_F$.

**Definition 5.1.** A function $f\colon \mathbb{N} \to (F[y])^*$ is called an $(\ell(n), n)$-*pseudo-random generator* against $\mathcal{A}_F$ if the following holds:

- $f(n) \in (F[y])^{n+1}$ for every $n > 0$.
- Let $f(n) = (f_1(y), \ldots, f_n(y), g(y))$. Then each $f_i(y)$ as well as $g(y)$ is a polynomial of degree at most $2^{\ell(n)}$.
- For any circuit $C \in \mathcal{A}_F$ of size $n$ with $m \leq n$ inputs:

$$C(x_1, x_2, \ldots, x_m) = 0 \text{ iff } C(f_1(y), f_2(y), \ldots, f_m(y)) = 0 \pmod{g(y)}.$$

A direct application of Schwartz–Zippel lemma [17], [26] shows that there always exist $(O(\log n), n)$-pseudo-random generators against $\mathcal{AC}_F$. Call such generators *optimal* pseudo-random generators. Pseudo-random generators that can be efficiently computed are of special interest.

**Definition 5.2.** An $(\ell(n), n)$-pseudo-random generator $f$ against $\mathcal{A}_F$ is *efficiently computable* if $f(n)$ is computable in time $2^{O(\ell(n))}$.

An easy argument shows that if there exists an efficiently computable $(\ell(n), n)$-pseudo-random generator against $\mathcal{AC}_F$ then the Identity Testing problem can be solved deterministically in time $2^{O(\ell(n))}$: evaluate the given circuit $C$ of size $n$ modulo $g(y)$ after substituting for the $i^{\text{th}}$ input variable the polynomial $f_i(y)$ where $f(n) = (f_1(y), \ldots, f_n(y), g(y))$. In particular, if there exists an efficiently computable optimal pseudo-random generator against $\mathcal{AC}_F$ then Identity Testing can be solved in polynomial time.

An efficiently computable pseudo-random generator also results in a lower bound.

**Theorem 5.3** ([1]). *Let $f$ be an efficiently computable $(\ell(n), n)$-pseudo-random generator against $\mathcal{A}_F$. Then there is a multilinear polynomial in $2\ell(n)$ variables, computable in time $2^{O(\ell(n))}$, that cannot be computed by any circuit in $\mathcal{A}_F$ of size $n$.*

*Proof.* For any $m = \ell(n)$, define the polynomial $q_f(x_1, x_2, \ldots, x_{2m})$ by

$$q_f(x_1, x_2, \ldots, x_{2m}) = \sum_{S \subseteq [1,2m]} c_S \cdot \prod_{i \in S} x_i.$$

The coefficients $c_S$ satisfy the condition

$$\sum_{S\subseteq[1,2m]} c_S \cdot \prod_{i\in S} f_i(y) = 0$$

where $f(n) = (f_1(y), f_2(y), \ldots, f_n(y), g(y))$. Such a $q_f$ always exists as the following argument shows.

> The number of coefficients of $q_f$ are exactly $2^{2m}$. These need to satisfy a polynomial equation of degree at most $2m \cdot 2^m$. So the equation gives rise to at most $2m \cdot 2^m + 1$ homogeneous constraints on the coefficients. Since $(2m \cdot 2^m + 1) < 2^{2m}$ for $m \geq 3$, there is always a non-trivial polynomial $q_f$ satisfying all the conditions.

The polynomial $q_f$ can be computed by solving a system of $2^{O(m)}$ linear equations in $2^{O(m)}$ variables over the field $F$. Each of these equations can be computed in time $2^{O(m)}$ using computability of $f$. Therefore, $q_f$ can be computed in time $2^{O(m)}$. Now suppose $q_f$ can be computed by a circuit $C \in \mathcal{A}_F$ of size $n$. By the definition of the polynomial $q_f$, it follows that $C(f_1(y), f_2(y), \ldots, f_{2m}(y)) = 0$. The size of circuit $C$ is $n$ and it computes a non-zero polynomial. This contradicts the pseudo-randomness of $f$. □

A partial converse of this theorem can also be shown: if there exists a polynomial family computable in time $2^{O(\ell(n))}$ that cannot be computed by any size $n$ circuit family in $\mathcal{A}_F$ then there exists an efficiently computable $(\ell^2(n), n)$-pseudo-random generator against $\mathcal{A}_F$, when the degree of every size $n$ circuit in $\mathcal{A}_F$ is bounded by $n^{O(1)}$.

An efficient optimal pseudo-random generator against $\mathcal{AC}_F$ yields a polynomial that requires exponential (in the number of variables) sized circuits. However, it is not clear whether the polynomial $q_f$ can be computed as permanent of a matrix of size $m^{O(1)}$. To get this, one needs to show that all the coefficients $c_S$ of $q_f$ are themselves efficiently computable. If this is done, then using the VNP characterization of permanent, it follows that $q_f$ equals the permanent of a matrix of size $m^{O(1)}$. This results in an exponential lower bound on permanent.

For a superpolynomial lower bound one needs either an $(n^{o(1)}, n)$-pseudo random generator against $\mathcal{AC}_F$ or an optimal pseudo-random generators against a much smaller class of circuits.

**Theorem 5.4** ([1]). *Let $f$ be an efficiently computable optimal pseudo-random generator against the class of circuits of depth $\omega(1)$ such that the associated polynomial $q_f$ is in* VNP. *Then permanent cannot computed by any polynomial sized circuit.*

*Proof.* From the previous theorem, it follows that the polynomial $q_f$ cannot be computed by exponential sized circuits of depth $\omega(1)$. A size $n^d$, depth $d \log n$ arithmetic circuit with fanin two multiplication gates can be translated to a subexponential sized

depth $d$ circuit by "cutting" the circuit into $\log n$ layers of depth $d$ each, and then "flattening" each layer to a subexponential sized circuit of depth two. Since every polynomial sized circuit computing permanent can be transformed to a depth $O(\log n)$, size $n^{O(1)}$ circuit with fanin two multiplication gates [23], the theorem follows. □

It is not clear at the moment how to construct optimal pseudo-random generators against constant depth circuits. In [1] a generator is conjectured. Unconditionally, we only know generators against depth two, polynomial sized circuits (the proof is easy, see [1]). We know an optimal generator against the following very special class of circuits too:

$$\mathcal{A} = \{C_n(x) \mid C_n(x) = (1+x)^n - 1 - x^n \text{ over the ring } \mathbb{Z}_n\}.$$

Notice that the circuits in the class $\mathcal{A}$ are not over a fixed field (or ring), and the size of the circuit $C_n$ is $O(\log n)$ and the degree is $n$. In [2], the following optimal generator was constructed against $\mathcal{A}$:

$$f(m) = \Big(x, 0, \ldots, 0, x^{16m^5} \cdot \prod_{r=1}^{16m^5} \prod_{a=1}^{4m^4} ((x-a)^r - 1)\Big).$$

## 6. Concluding remarks

The problem of proving that the permanent of a size $n$ matrix cannot be expressed as determinant of size $n^{O(\log n)}$ matrix is of great importance in complexity theory. While the existing approaches have failed to shed light on this, one hopes that at least one of the two new approaches will eventually lead to a solution of the problem.

**Acknowledgements.** I wish to thank Somenath Biswas for enjoyable discussions and help in preparing this article.

## References

[1] Agrawal, M., Proving lower bounds via pseudo-random generators. In *Foundations of Software Technology and Theoretical Computer Science*, Lecture Notes in Comput. Sci. 3821, Springer-Verlag, Berlin 2005, 92–105.

[2] Agrawal, M., On derandomizing tests for certain polynomial identities. In *Proceedings of 18th Annual IEEE Conference on Computational Complexity*, IEEE Computer Society, Los Alamitos, CA, 2003, 355–362.

[3] Babai, L., Fortnow, L., Nisan, N., and Wigderson, A., BPP has subexponential time simulations unless EXPTIME has publishable proofs. *Comput. Complexity* **3** (4) (1963), 307–318.

[4] Damm, C., DET=$\mathrm{L}^{\#L}$. Technical Report Informatik, Preprint 8, Fachbereich Informatik der Humboldt Universität zu Berlin, 1991.

[5] Goldreich, O., *Foundation of Cryptography I: Basic Tools*. Cambridge University Press, Cambridge 2001.

[6] Grigoriev, D., and Razborov, A:, Exponential lower bounds for depth 3 arithmetic circuits in algebras of functions over finite fields. *Appl. Algebra Engrg. Comm. Comput.* **10** (6) (2000), 467–487, 2000.

[7] Impagliazzo, R., and Wigderson, A., P = BPP if E requires exponential circuits: Derandomizing the XOR lemma. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1997, 220–229.

[8] Jerrum, M., and Snir, M., Some exact complexity results for straight-line computations over semirings. *J. ACM* **29** (3) (1982), 874–897.

[9] Kabanets, Valentine, and Impagliazzo, Russell, Derandomizing polyonmial identity tests means proving circuit lower bounds. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2003, 355–364.

[10] Minc, H., *Permanents*. Addision-Wesley, 1978.

[11] Mulmulay, K., and Sohoni, M., Geometric complexity theory, P vs. NP, and explicit obstructions. *SIAM J. Comput.* **31** (2) (2002), 496–526.

[12] D. Mumford, D., *Algebraic Geometry I: Complex Projective Varieties*. Grundlehren Math. Wiss. 221, Springer-Verlag, Berlin 1976.

[13] Nisan, N., and Wigderson, A., Hardness vs. randomness. *J. Comput. System Sci.* **49** (2) (1994), 149–167.

[14] Nisan, N., and Wigderson, A., Lower bounds on arithmetic circuits via partial derivatives. *Comput. Complexity* **6** (3) (1996/97), 217–234.

[15] Raz, Ran, Multi-linear formulas for permanent and determinant and of super-polynomial size. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2004, 633–641.

[16] Regan, K., Understanding the Mulmuley-Sohoni approach to P vs. NP. *Bulletin of the European Association for Theoretical Computer Science* **78** (2002), 86–97. Lance Fortnow's Computational Complexity Column.

[17] Schwartz, J. T., Fast probabilistic algorithms for verification of polynomial identities. *J. ACM* **27** (4) (1980), 701–717.

[18] Seiferas, J., Fischer, M., and Meyer, A., Separating nondeterministic time complexity classes. *J. ACM* **25** (1) (1987), 146–167.

[19] Shpilka, A., and Wigderson, A., Depth-3 arithmetic circuits over fields of characteristic zero. *Comput. Complexity* **10** (1) (2001), 1–27.

[20] Sudan, M., Trevisan, L., and Vadhan, S., Pseudorandom generators without the XOR lemma. In *Proceedings of the 31th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1999, 537–546.

[21] Toda, S., Counting problems computationally equivalent to the determinant. Manuscript, 1991.

[22] Toda, S., PP is as hard as the polyonmial-time hierarchy. *SIAM J. Comput.* **20** (1991), 865–877.

[23] Valiant, L., Skyum, S., Berkowitz, S., and Rackoff, C., Fast parallel computation of polynnomials using few processors. *SIAM J. Comput.* **12** (1983), 641–644.

[24] Valiant, L., Completeness classes in algebra. In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1979, 249–261.

[25] Vinay, V., Counting auxiliary pushdown automata and semi-unbounded arithmetic circuits. In *Proceedings of the Structure in Complexity Theory Conference*, Lecture Notes in Comput. Sci. 223, Springer-Verlag, Berlin 1991, 270–284.

[26] Zippel, R. E., Probabilistic algorithms for sparse polynomials. In *EUROSCAM'79*, Lecture Notes in Comput. Sci. 72, Springer-Verlag, Berlin 1979, 216–226.

Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur 208016, India
E-mail: manindra@iitk.ac.in

# The additivity problem in quantum information theory

Alexander S. Holevo*

**Abstract.** In this lecture we survey the present status of the additivity problem for the classical capacity and related characteristics of quantum channels – one of the most profound mathematical problems of quantum information theory.

**Mathematics Subject Classification (2000).** Primary 94A15; Secondary 81P68.

**Keywords.** Quantum information theory, quantum channel, classical capacity, entanglement, additivity problem.

## 1. Introduction

The problems of data transmission and storage by quantum information carriers received increasing attention during past decade, owing to the burst of activity in the field of quantum information and computation [42], [22]. At present we are witnessing emergence of theoretical and experimental foundations of the quantum information science. It represents a new exciting research field addressing a number of fundamental issues both in quantum physics and in information and computer sciences. On the other hand, it provides a rich source of well-motivated mathematical problems, often having simple formulations but hard solutions.

A central result in the classical information theory is the *coding theorem*, establishing the possibility of reliable data transmission and processing at rates lower than the *capacity* of the communication channel. The issue of the information capacity of *quantum channels* arose soon after publication of Shannon's pioneering paper and goes back to the works of Gabor, Brillouin and Gordon, asking for fundamental limits on the rate and quality of information transmission. These works laid a physical foundation and raised the question of consistent mathematical treatment of the problem. Important steps in this direction were made in the seventies when quantum statistical decision theory was created, making a noncommutative probability frame for this circle of problems, see [21] for a survey.

A dramatic progress has been achieved during the past decade [42], [6], [22]. In particular, a number of coding theorems was discovered, moreover, it was realized

*This work was done partially when the author was Leverhulme Visiting Professor at CQC, DAMTP, University of Cambridge. The work was also supported by the Program "Theoretical Problems of Modern Mathematics" of the Mathematics Division of RAS. The author is grateful to A. Ekert, M. B. Ruskai, M. Shirokov, Yu. M. Suhov and R. F. Werner for fruitful discussions.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

that the quantum channel is characterized by the whole spectrum of capacities depending on the nature of the information resources and the specific protocols used for the transmission, see [6], [14]. This new age of quantum information science is characterized by emphasis onto the new possibilities (rather than mere restrictions) inherent in the quantum nature of the information processing agent. On the other hand, the questions of information capacities turned out to be relevant to the theory of quantum computations, particularly in connection with quantum error-correction, communication protocols, algorithmic complexity and a number of other important issues.

The quantum information processing systems have a specifically novel resource, *entanglement*, a kind of non-classical correlation between parts of the composite quantum system. Among many other unusual features it underlies the *strict superadditivity* of Shannon information due to *entangled decodings* in a situation formally similar to the classical memoryless channels [20], [22], [6]. Namely, for independent quantum systems there are entangled measurements which can bear more information than the arithmetic sum of information from these systems. This property has profound consequences for the theory of quantum communication channels and their capacities.

A closely related issue is the additivity of the capacity-related quantities for the memoryless quantum channels with respect to *entangled encodings*. Should the additivity fail, this would mean that applying entangled inputs to several independent uses of a quantum channel may result in superadditive increase of its capacity for transmission of classical information. However so far there is neither a single evidence of such a non-additivity, nor a general proof for the additivity. In this lecture we survey the present status of this problem.

We start in Section 2 with the classical case, where the additivity holds for almost obvious reasons. We then describe the problem in the finite-dimensional quantum setting in Section 3, discussing also the various formulations of the additivity conjecture and connections between them. Positive results for several concrete classes of channels are briefly surveyed in Section 4, where also an important counterexample to the additivity of the minimal output quantum Rényi entropy is discussed. Since quantum communication channels are described mathematically as completely positive maps, we devote Subsection 4.1 to the description of their structure paying attention to the notion of complementary maps which leads to new examples of additivity. Section 5 is devoted to different formulations of the additivity conjecture using tools from convex analysis. In Subsection 5.4 we present an argument, essentially due to P. Shor, implying the global equivalence of different forms of the additivity conjecture. We conclude with Section 6, where we briefly outline the works treating the infinite-dimensional case.

## 2. Additivity in the classical information theory

Let $\mathcal{X}$, $\mathcal{Y}$ be two finite sets (alphabets), and let $[\Phi(x, y)]_{x\in\mathcal{X}, y\in\mathcal{Y}}$ be a *stochastic matrix*, i.e.,

1. $\Phi(x, y) \geq 0,\ x \in \mathcal{X}, y \in \mathcal{Y}$;
2. $\sum_{y\in\mathcal{Y}} \Phi(x, y) = 1,\ x \in \mathcal{X}$.

In information theory a stochastic matrix describes a (noisy) *channel* from $\mathcal{X}$ to $\mathcal{Y}$. It transforms an input probability distribution $\pi$ on $\mathcal{X}$ into the output probability distribution $\pi' = \Phi\pi$ on $\mathcal{Y}$. Denote by

$$\mathcal{P}(\mathcal{X}) = \left\{\pi : \pi(x) \geq 0,\ \sum_{x\in\mathcal{X}} \pi(x) = 1\right\}$$

the simplex of all probability distributions $\pi$ on $\mathcal{X}$. Extreme points of $\mathcal{P}(\mathcal{X})$ are the degenerate probability distributions $\delta_x$ on $\mathcal{X}$. Notice the following obvious property:

*For a direct product $\mathcal{X}_1 \times \mathcal{X}_2$ of two alphabets, extreme points of $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ are precisely the products of extreme points of $\mathcal{P}(\mathcal{X}_j)$*:

$$\mathrm{ext}\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathrm{ext}\mathcal{P}(\mathcal{X}_1) \times \mathrm{ext}\mathcal{P}(\mathcal{X}_2). \tag{1}$$

The most important characteristic of a channel is its *capacity*

$$C(\Phi) = \max_{\pi\in\mathcal{P}(\mathcal{X})} \left\{H(\Phi\pi) - \sum_x \pi(x) H(\Phi\delta_x)\right\}, \tag{2}$$

where the expression in curly brackets is equal to the Shannon *mutual information* between the input and the output of the channel. Here

$$H(\pi) = -\sum_x \pi(x) \log \pi(x)$$

is the entropy of the probability distribution $\pi$.

One of the main results of information theory – the coding theorem for memoryless channels, see e.g. [10] – says that the quantity (2) is the ultimate rate of asymptotically perfect transmission of information by $n$ independent uses of the channel $\Phi$, when $n \to \infty$. The capacity has the fundamental additivity property

$$C(\Phi_1 \otimes \Phi_2) = C(\Phi_1) + C(\Phi_2). \tag{3}$$

Here the inequality $\geq$ (superadditivity) follows by restricting to the independent inputs, while the opposite inequality can be proved by using subadditivity of the output entropy $H(\Phi\pi)$ and the property (1) for the second term in the Shannon information (which is equal to minus the conditional output entropy). The additivity is an important ingredient of the proof of the coding theorem, implying

$$C(\Phi^{\otimes n}) = nC(\Phi),$$

where $\Phi^{\otimes n} = \underbrace{\Phi \otimes \cdots \otimes \Phi}_{n}$. It expresses the "memoryless" character of the information transmission scheme based on the independent uses of the channel. For schemes with memory the capacity can be strictly superadditive.

In what follows we are going to describe the noncommutative analog of the quantity $C(\Phi)$, as well as several other related quantities playing a basic role in quantum information theory. The corresponding additivity property was conjectured to hold also in the noncommutative case, although so far there is neither a general proof, nor a counterexample; moreover the additivity is no longer "natural" since an analog of the underlying basic fact (1) breaks dramatically in the noncommutative case.

## 3. Quantum channels

**3.1. The $\chi$-capacity.** Let $\mathcal{H}$ be a unitary space and let $\mathfrak{M}(\mathcal{H})$ be the algebra of all linear operators in $\mathcal{H}$. By choosing an orthonormal basis, $\mathcal{H}$ can be identified with the space $\mathcal{H}_d$ of $d$-dimensional complex vectors and $\mathfrak{M}(\mathcal{H})$ with the algebra $\mathfrak{M}_d$ of complex $d \times d$-matrices.

We shall consider linear maps $\Phi$ which take operators $F$ in $d$-dimensional unitary space $\mathcal{H}$ to operators $F' = \Phi(F)$ in a $d'$-dimensional space $\mathcal{H}'$. Sometimes these are called "superoperators" or "supermatrices" because they can be described as matrices with $d^2 \times d'^2$ entries [9].

Let $\Phi_j \colon \mathfrak{M}(\mathcal{H}_j) \to \mathfrak{M}(\mathcal{H}'_j)$; $j = 1, 2$, be two such maps, and let $\Phi_1 \otimes \Phi_2 : \mathfrak{M}(\mathcal{H}_1) \otimes \mathfrak{M}(\mathcal{H}_2) \to \mathfrak{M}(\mathcal{H}'_1) \otimes \mathfrak{M}(\mathcal{H}'_2)$ be their tensor product defined by the natural action on product operators and then extended by linearity.

An operator $F \in \mathfrak{M}(\mathcal{H})$ is called positive, $F \geq 0$, if the corresponding matrix is positive semidefinite and the map $\Phi \colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ is called *positive* if $F \geq 0$ implies $\Phi(F) \geq 0$.

Especially important for us will be the class of completely positive (CP) maps [51], [9], [43]. The map $\Phi \colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ is *completely positive*, if for $d = 1, 2, \ldots$ the maps $\Phi \otimes \mathrm{Id}_d$ are all positive, where $\mathrm{Id}_d \colon \mathfrak{M}_d \to \mathfrak{M}_d$ is the identity map of the algebra of $d \times d$- matrices. It follows that the tensor product of CP maps is again CP, since

$$\Phi_1 \otimes \Phi_2 = (\mathrm{Id}_{d'_1} \otimes \Phi_2) \circ (\Phi_1 \otimes \mathrm{Id}_{d_2}).$$

There are positive maps that are not CP, a basic example being provided by the matrix transposition $F \to F^\top$ in a fixed basis.

Finite quantum system is described by a unitary space $\mathcal{H}$. The convex subset

$$\mathfrak{S}(\mathcal{H}) = \{\rho : \rho \geq 0,\ \mathrm{Tr}\,\rho = 1\}$$

of $\mathfrak{M}(\mathcal{H})$ is called the *quantum state space*. The operators $\rho \in \mathfrak{S}(\mathcal{H})$ are called *density operators* or *quantum states*. The state space is a compact convex set with the

extreme boundary

$$\mathfrak{P}(\mathcal{H}) = \mathrm{ext}\mathfrak{S}(\mathcal{H}) = \{\rho : \rho \geq 0,\ \mathrm{Tr}\,\rho = 1,\ \rho^2 = \rho\}.$$

Thus extreme points of $\mathfrak{S}(\mathcal{H})$, which are also called *pure states*, are one-dimensional projectors, $\rho = P_\psi$ for a vector $\psi \in \mathcal{H}$ with unit norm, see, e.g. [42], [22].

Instead of the classical relation (1), one has the following relation for a tensor product $\mathcal{H}_1 \otimes \mathcal{H}_2$ of two unitary spaces

$$\mathrm{ext}\mathfrak{S}(\mathcal{H}_1 \otimes \mathcal{H}_2) \supsetneq \mathrm{ext}\mathfrak{S}(\mathcal{H}_1) \times \mathrm{ext}\mathfrak{S}(\mathcal{H}_2), \tag{4}$$

since apparently there are continually many pure states $P_\psi$ in $\mathcal{H}_1 \otimes \mathcal{H}_2$, given by vectors $\psi$ not representable as a tensor product $\psi_1 \otimes \psi_2$. In quantum theory the tensor product $\mathcal{H}_1 \otimes \mathcal{H}_2$ describes the composite (bipartite) system. Vectors that are not of the form $\psi_1 \otimes \psi_2$, as well as the corresponding pure states, are called *entangled*. In an entangled pure state of a bipartite quantum system, neither of the parts is in a pure state, in a sharp contrast to the classical systems.

A CP map $\Phi$ is called a (quantum) *channel*, if it is trace preserving, i.e. if it maps quantum states into quantum states (possibly in another unitary space $\mathcal{H}'$). A channel $\Phi$ is called *unital* if $d = d'$ and $\Phi(I) = I'$, where $I(I')$ is the identity operator in $\mathcal{H}$ resp. $\mathcal{H}'$.

The von Neumann entropy of a density operator $\rho$

$$H(\rho) = -\,\mathrm{Tr}\,\rho \log \rho$$

is nonnegative concave continuous function on $\mathfrak{S}(\mathcal{H})$ vanishing on $\mathfrak{P}(\mathcal{H})$ and taking the maximal value $\log d$ on the *chaotic state* $\rho = \frac{I}{d}$. The noncommutative analog of the quantity (2) is the $\chi$*-capacity* [20], [22] of the channel $\Phi$,

$$C_\chi(\Phi) = \max_\pi \left\{ H\Big(\Phi\Big(\sum_x \pi(x)\rho(x)\Big)\Big) - \sum_x \pi(x) H(\Phi(\rho(x))) \right\}, \tag{5}$$

where the maximum is taken over all *state ensembles* i.e. finite probability distributions $\pi$ on the quantum state space $\mathfrak{S}(\mathcal{H})$ ascribing probabilities $\pi(x)$ to density operators $\rho(x)$[1]. The *additivity conjecture* is whether the analog of the property (3) holds for quantum channels, i.e.

$$C_\chi(\Phi_1 \otimes \Phi_2) \stackrel{?}{=} C_\chi(\Phi_1) + C_\chi(\Phi_2). \tag{6}$$

Here again $\otimes$ is the tensor product of the two channels describing independent uses of the channels on the states of the composite system. This is the earliest additivity conjecture in quantum information theory which can be traced back to [5], see also [20], [6].

[1] In the finite dimensional case we are considering the maximum is indeed attained on $\pi$ with support having at most $d^2$ states, where $d = \dim\mathcal{H}$ [50].

In physical terms this problem can also be formulated as: "Can entanglement between input states help to send classical information through quantum channels?" The *classical capacity* of a quantum channel is defined as the maximal transmission rate per use of the channel, with coding and decoding chosen for an increasing number $n$ of independent uses of the channel

$$\Phi^{\otimes n} = \underbrace{\Phi \otimes \cdots \otimes \Phi}_{n}$$

such that the error probability goes to zero as $n \to \infty$. A basic result of quantum information theory – the quantum coding theorem [19], [49] – implies that the classical capacity $C(\Phi)$ and the $\chi$-capacity $C_\chi(\Phi)$ are connected by the formula

$$C(\Phi) = \lim_{n\to\infty}(1/n)C_\chi(\Phi^{\otimes n}).$$

Since $C_\chi(\Phi)$ is easily seen to be superadditive, i.e.

$$C_\chi(\Phi_1 \otimes \Phi_2) \geq C_\chi(\Phi_1) + C_\chi(\Phi_2),$$

one has $C(\Phi) \geq C_\chi(\Phi)$. If the additivity (6) holds, then $C_\chi(\Phi^{\otimes n}) = nC_\chi(\Phi)$, and this would imply $C(\Phi) = C_\chi(\Phi)$. Such a result would be very much welcome from a mathematical point of view, giving a relatively easily computable "single-letter" expression for the classical capacity of a quantum channel.

On the other hand, such an equality is rather counter-intuitive in view of the relation (4) and existence of waste variety of pure entangled states. In fact, there are several quantities that are nonadditive under the tensor product of quantum channels such as: a) the Shannon information maximized over entangled outputs [20]; b) the quantum capacity [6]; c) the minimal output Rényi entropy [53] and some others, the classical counterparts of which are additive. In the following we shall consider the case c) which is most relevant to our main problem (6).

**3.2. Entropic characteristics of CP maps and channels.** The quantum Rényi entropy of order $p > 1$ of a density operator $\rho$ is defined as

$$R_p(\rho) = \frac{1}{1-p}\log \operatorname{Tr} \rho^p, \tag{7}$$

so that the minimal output Rényi entropy of the channel $\Phi$ is

$$\check{R}_p(\Phi) = \min_{\rho\in\mathfrak{S}(\mathcal{H})} R_p(\Phi(\rho)) = \frac{p}{1-p}\log \nu_p(\Phi),$$

where

$$\nu_p(\Phi) = \max_{\rho\in\mathfrak{S}(\mathcal{H})} [\operatorname{Tr} \Phi(\rho)^p]^{1/p} \tag{8}$$

is a "measure of output purity" of the channel $\Phi$ introduced in [3][2]. In the limit $p \downarrow 1$ the quantum Rényi entropies monotonically increase and uniformly converge to the entropy of a density operator $\rho$,

$$\lim_{p\downarrow 1} R_p(\rho) = H(\rho),$$

so that introducing the minimal output entropy

$$\check{H}(\Phi) = \min_{\rho\in\mathfrak{S}(\mathcal{H})} H(\Phi(\rho)) \tag{9}$$

of the quantum channel $\Phi$, one has $\lim_{p\downarrow 1} \check{R}_p(\Phi) = \check{H}(\Phi)$.

The classical analog of the quantity (8) is

$$\nu_p(\Phi) = \max_{\pi\in\mathcal{P}(\mathcal{X})} \|\Phi\pi\|_p,$$

where $\|f\|_p = \left(\sum_{x\in\mathcal{X}} |f(x)|^p\right)^{1/p}$ is the $l_p$-norm of $f = (f(x))_{x\in\mathcal{X}}$. The function $\pi \to \|\Phi\pi\|_p$ is convex continuous and hence attains the maximum at an extreme point of $\mathcal{P}(\mathcal{X})$, i.e. on a degenerate probability distribution $\delta_x$. Hence the basic property (1) implies the multiplicativity relation

$$\nu_p(\Phi_1 \otimes \Phi_2) = \nu_p(\Phi_1)\nu_p(\Phi_2), \tag{10}$$

which is equivalent to the additivity property of the minimal output Rényi entropies

$$\check{R}_p(\Phi_1 \otimes \Phi_2) = \check{R}_p(\Phi_1) + \check{R}_p(\Phi_2), \tag{11}$$

implying in turn

$$\check{H}(\Phi_1 \otimes \Phi_2) = \check{H}(\Phi_1) + \check{H}(\Phi_2), \tag{12}$$

in the limit $p \downarrow 1$. Notice that the inequality $\leq$ is obvious in (11), (12).

Unlike the classical case, there is no apparent reason for these multiplicativity/additivity properties to hold in the case of quantum channels. Nevertheless there are several important classes of channels for which the multiplicativity (10) can be proved for all $p > 1$, although there is also an example where it breaks for sufficiently large $p$. This, however, does not preclude that it can hold for $p$ close to 1, and the validity of (10) for $p \in (1, 1+\varepsilon)$, with $\varepsilon > 0$, implies validity of the additivity property (12), which, as we shall see, is closely related to the additivity of the $\chi$-capacity (6).

Here we would also like to mention that multiplicativity of more general $(q \to p)$-norms was studied for the cases where at least some of the maps $\Phi_1$, $\Phi_2$ is not CP, see [38], [37], [35]. Basing on the advanced theory of the operator $L_p$-spaces [44], [43], there is an interesting study concerning the multiplicativity of *completely bounded* $p$-norms, which however is related to the additivity of a completely different entropic quantity [15].

[2] In the finite-dimensional case all the functions of the state we are considering are easily seen to be continuous and their extrema on the state space are attained. However it is not so in infinite-dimensional case, and then the attainability of the extrema requires separate study, see Section 6.

## 4. Some classes of CP maps and channels

**4.1. Representations of CP maps.** Here we recollect some facts concerning the structure of CP maps and channels. Given three unitary spaces $\mathcal{H}_A$, $\mathcal{H}_B$, $\mathcal{H}_C$ and a linear operator $V\colon \mathcal{H}_A \to \mathcal{H}_B \otimes \mathcal{H}_C$, the relation

$$\Phi(\rho) = \mathrm{Tr}_{\mathcal{H}_C} V\rho V^*, \quad \tilde{\Phi}(\rho) = \mathrm{Tr}_{\mathcal{H}_B} V\rho V^*; \quad \rho \in \mathfrak{M}(\mathcal{H}_A) \tag{13}$$

defines two CP maps $\Phi\colon \mathfrak{M}(\mathcal{H}_A) \to \mathfrak{M}(\mathcal{H}_B)$, $\tilde{\Phi}\colon \mathfrak{M}(\mathcal{H}_A) \to \mathfrak{M}(\mathcal{H}_C)$, which are called mutually *complementary* [25] (or conjugate [36]). If $V$ is an isometry then both maps are channels.

For any linear map $\Phi\colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ the *dual* map $\Phi^*\colon \mathfrak{M}(\mathcal{H}') \to \mathfrak{M}(\mathcal{H})$ is defined by the formula

$$\mathrm{Tr}\,\Phi(\rho)X = \mathrm{Tr}\,\rho\Phi^*(X); \quad \rho \in \mathfrak{M}(\mathcal{H}),\ X \in \mathfrak{M}(\mathcal{H}').$$

If $\Phi$ is CP, then $\Phi^*$ is also CP. The relations (13) are equivalent to

$$\Phi^*(X) = V^*(X \otimes I_C)V; \quad X \in \mathfrak{M}(\mathcal{H}_B), \tag{14}$$

$$\tilde{\Phi}^*(X) = V^*(I_B \otimes X)V; \quad X \in \mathfrak{M}(\mathcal{H}_C). \tag{15}$$

The *Stinespring dilation theorem* [51] concerning CP maps on arbitrary $C^*$-algebras, for the particular case in question amounts to the statement that for a given CP map there are a space $\mathcal{H}_C$ and an operator $V$ satisfying (14). This implies that given a CP map $\Phi$, a complementary map $\tilde{\Phi}$ always exists.

By introducing a basis $\{e_j^C\}$ in $\mathcal{H}_C$ and operators $V_j\colon \mathcal{H}_A \to \mathcal{H}_B$ defined by

$$(\varphi, V_j\psi) = \left(\varphi \otimes e_j^C, V\psi\right); \quad \varphi \in \mathcal{H}_B,\ \psi \in \mathcal{H}_A,$$

the first relation in (13) can be rewritten as

$$\Phi(\rho) = \sum_{j=1}^{d_C} V_j \rho V_j^*; \quad \rho \in \mathfrak{M}(\mathcal{H}_A). \tag{16}$$

The map (16) is a channel if and only if $\sum_{j=1}^{d_C} V_j^* V_j = I$. The relation (16) is usually called the Kraus representation (see also Choi [9]). Of course, there are similar representations for the complementary map $\tilde{\Phi}$ and the dual maps.

**Theorem 4.1** ([25], [36]). *If one of the relations* (11), (12) *holds for the CP maps* (*channels*) $\Phi_1$, $\Phi_2$, *then similar relations holds for the pair of their complementary maps* $\tilde{\Phi}_1$, $\tilde{\Phi}_2$. *If one of these relations holds for a given* $\Phi_1$ *and arbitrary* $\Phi_2$, *then a similar relation holds for the complementary map* $\tilde{\Phi}_1$ *and arbitrary* $\Phi_2$.

Validity of the multiplicativity conjecture (10) for all $p \geq 1$ and of the additivity conjectures (12), (6) was established in a number of cases where one channel is arbitrary and the other belongs to one of the classes we are going to discuss.

**4.2. Entanglement-breaking maps and their complementary maps.** Any linear map $\Phi\colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ admits a representation

$$\Phi(X) = \sum_j N'_j \operatorname{Tr} X M_j, \tag{17}$$

where $\{M_j\}$, $\{N'_j\}$ are finite collections of operators in $\mathcal{H}$ and $\mathcal{H}'$, respectively. This simply follows from the finite dimensionality of $\mathcal{H}$, $\mathcal{H}'$ and the fact that any linear functional on $\mathfrak{M}(\mathcal{H})$ has the form $X \to \operatorname{Tr} XM$, where $M \in \mathfrak{M}(\mathcal{H})$.

**Proposition 4.2.** *For a linear map $\Phi\colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ the following conditions are equivalent:*

(i) *There is a representation* (17) *such that $M_j \geq 0$, $N'_j \geq 0$.*

(ii) *The map $\Phi$ is CP and has the representation* (16) *with rank one operators $V_j$.*

(iii) *For $d = 2, 3, \ldots$ and any $\rho_{12} \in \mathfrak{S}(\mathcal{H} \otimes \mathcal{H}_d)$,*

$$(\Phi \otimes \mathrm{Id}_d)(\rho_{12}) = \sum_\alpha A_\alpha \otimes B_\alpha \tag{18}$$

*where $A_\alpha \geq 0$ and $B_\alpha \geq 0$.*

Channels satisfying the condition (i) were introduced in [20], and the above characterization was obtained in [29] where such maps were termed *entanglement-breaking*. In the case of channels, (18) means that the output state $(\Phi \otimes \mathrm{Id}_d)(\rho_{12})$ is always *separable*, i.e. a convex combination of (unentangled) product states. Entanglement-breaking channels can be written in the form

$$\Phi(\rho) = \sum_j \rho'_j \operatorname{Tr} \rho M_j, \tag{19}$$

where $\{\rho'_j\}$ is a finite collection of states in $\mathcal{H}'$, and $\{M_j\}$ a *resolution of the identity* in $\mathcal{H}$, i.e. a collection of operators satisfying

$$M_j \geq 0, \quad \sum_j M_j = I.$$

Resolutions of the identity describe quantum *observables* [22], and the channel (19) corresponds to a measurement of the observable $\{M_j\}$ over an input state $\rho$ resulting in a probability distribution $\{\operatorname{Tr} \rho M_j\}$, which is followed by preparation of the output state $\rho'_j$. Thus, there is a classical information processing stage inside the channel which is responsible for the entanglement-breaking.

The simplest example is the *completely depolarizing channel*

$$\Phi(\rho) = \frac{I}{d} \operatorname{Tr} \rho$$

which maps an arbitrary state to the chaotic state $\frac{I}{d}$.

As shown in [25], [36], the complementary maps to the entanglement-breaking maps have the form

$$\tilde{\Phi}(\rho) = \sum_{j,k=1}^{d_C} c_{jk}\langle\psi_j|\rho|\psi_k\rangle E_{jk}, \quad \rho \in \mathfrak{M}(\mathcal{H}_A), \tag{20}$$

where $[c_{jk}]$ is a nonnegative definite matrix, $\{\psi_j\}_{j=\overline{1,d_C}}$ a system of vectors in $\mathcal{H}_A$, and the $E_{jk}$'s are the matrix units in $\mathcal{H}_C$. In the special case where $\{\psi_j\}_{j=\overline{1,d_C}}$ is an orthonormal basis, (20) is the *diagonal* CP map in the sense of [31]. Diagonal channels are characterized by the additional property $c_{jj} \equiv 1$.

A simplest example of the diagonal channel is the *ideal channel* Id, which is complementary to the completely depolarizing channel.

For general entanglement-breaking channels the additivity property (12) with arbitrary second channel was established by Shor [47], preceded by results in [20] on special subclasses of such channels. The multiplicativity property (10) for all $p > 1$ was established by King [32], basing on the Lieb–Thirring inequality [40]: for $A, B \in \mathfrak{M}(\mathcal{H})$, $A, B \geq 0$, and $p \geq 1$

$$\operatorname{Tr}(AB)^p \leq \operatorname{Tr} A^p B^p. \tag{21}$$

By Theorem 4.1 this implies the corresponding properties for the complementary maps and channels of the form (20).

**4.3. Covariant channels.** Let $G$ be a group (either finite or continuous) and let $g \to U_g^A, U_g^B$, $g \in G$, be two projective (unitary) representations of $G$ in $\mathcal{H}_A$, $\mathcal{H}_B$. The CP map $\Phi\colon \mathfrak{M}(\mathcal{H}_A) \to \mathfrak{M}(\mathcal{H}_B)$ is *covariant* if

$$\Phi(U_g^A \rho U_g^{A*}) = U_g^B \Phi(\rho) U_g^{B*} \tag{22}$$

for all $g \in G$ and all $\rho$. For a covariant CP map there exists a covariant Stinespring dilation: namely, there is a projective representation $g \to U_g^C$ in $\mathcal{H}_B$, such that

$$(U_g^B \otimes U_g^C)V = V U_g^A,$$

see e.g. [24]. It follows that the complementary map is also covariant:

$$\tilde{\Phi}(U_g^A \rho U_g^{A*}) = U_g^C \tilde{\Phi}(\rho) U_g^{C*}.$$

**Lemma 4.3.** *If the representation $U_g^A$ is irreducible, then*

$$C_\chi(\Phi) = H\left(\Phi\left(\frac{I_A}{d_A}\right)\right) - \check{H}(\Phi). \tag{23}$$

Since the tensor product of irreducible representations of possibly different groups $G_1$, $G_2$ is an irreducible representation of the group $G_1 \times G_2$, it follows that the additivity properties (12) and (6) are equivalent for channels satisfying the condition of Lemma 4.3. Symmetry considerations also help to compute explicitly the entropic characteristics of covariant channels. Then, in the case of additivity, $C = C_\chi$ gives an explicit expression for the classical capacity of the channel.

**4.4. The unital qubit channels.** The simplest and yet fundamental quantum system is the *qubit* (quantum bit), where $\dim\mathcal{H} = 2$. A convenient basis in $\mathfrak{M}_2$ is formed by the Pauli matrices

$$\sigma_0 \equiv I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

It is known [45] that an arbitrary unital channel $\Phi\colon \mathfrak{M}_2 \to \mathfrak{M}_2$ can be decomposed as

$$\Phi(\rho) = U_2\Lambda(U_1\rho U_1^*)U_2^*, \tag{24}$$

where $U_1$, $U_2$ are unitary matrices and $\Lambda$ has the following canonical Kraus representation:

$$\Lambda(\rho) = \sum_{\gamma=0,x,y,z} \mu_\gamma \sigma_\gamma \rho \sigma_\gamma, \tag{25}$$

where $\{\mu_\gamma\}$ is a probability distribution. The unital qubit channels (25) are covariant with respect to the projective representation of the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ defined by

$$U_{00} = \sigma_0, \quad U_{01} = \sigma_z, \quad U_{10} = \sigma_x, \quad U_{11} = -i\sigma_y.$$

Therefore the relation (23) holds for this class of channels.

By using a convex decomposition into diagonal channels of special form and applying to these the Lieb–Thirring inequality (21), King [33] established (10) for all $p > 1$, (12) and (6) for the case where $\Phi_1$ is an arbitrary unital qubit channel and $\Phi_2$ is an arbitrary channel. There are recent positive results concerning nonunital qubit channels [35].

**4.5. Depolarizing channel.** The *depolarizing channel* in $\mathcal{H}_d$ is

$$\Phi(\rho) = (1-p)\rho + p\frac{I}{d}\operatorname{Tr}\rho, \quad 0 \le p \le \frac{d^2}{d^2-1}. \tag{26}$$

If $p \le 1$ this describes a mixture of the ideal channel Id and the completely depolarizing channel. For the whole range $0 \le p \le \frac{d^2}{d^2-1}$ complete positivity can be proven by using the Kraus decomposition, see e.g. [42]. The depolarizing channel is characterized by the property of unitary covariance,

$$\Phi(U\rho U^*) = U\Phi(\rho)U^*,$$

for an arbitrary unitary operator $U$ in $\mathcal{H}$.

The properties (10) for all $p > 1$, (12) and (6) were proved in [34] for the case where $\Phi_1$ is a depolarizing channel and $\Phi_2$ is arbitrary, using a method similar to the case of the unital qubit channels.

Complementarity for depolarizing channels is computed in [11].

**4.6. A transpose-depolarizing channel.** Let us consider in some detail the *extreme transpose-depolarizing* channel

$$\Phi(\rho) = \frac{1}{d-1}\big[I \operatorname{Tr} \rho - \rho^{\top}\big], \tag{27}$$

where $\rho^{\top}$ is the transpose of $\rho$ in an orthonormal basis $\{e_j\}$ in $\mathcal{H}_d$. Complete positivity of the map (27) follows from the representation

$$\Phi(\rho) = \frac{1}{2(d-1)} \sum_{j,k=1}^{d} (E_{jk} - E_{kj})\rho(E_{jk} - E_{kj})^*. \tag{28}$$

It has the covariance property

$$\Phi(U\rho U^*) = \overline{U}\Phi(\rho)\overline{U}^*$$

for an arbitrary unitary $U$, where $\overline{U}$ is complex conjugate in the basis $\{e_j\}$. It follows that the relation (23) holds for this channel.

This channel is interesting in that it breaks the additivity of the minimal Rényi entropy (11) with $\Phi_1 = \Phi_2 = \Phi$ for $d > 3$ and large enough $p$ [53]. At the same time it fulfills (12), see [41], [12], and even (11) for $1 \le p \le 2$ [13]. For generalizations to broader classes of channels as well as to the more general forms of additivity, see [41], [1], [54]. This example also shows that although the Lieb–Thirring inequality can be used in several cases to prove the additivity conjecture (11) for all $p > 1$, it cannot serve for a general proof. Moreover, there is even no general proof covering all these cases, since each time application of the Lieb–Thirring inequality is supplied with an argument specific to the case under consideration.

The complementary channel which shares the multiplicativity/additivity properties with the channel (27) is

$$\tilde{\Phi}(\rho) = \frac{2}{(d-1)} P_-(\rho \otimes I_2) P_- \tag{29}$$

(see [25] for more details). Here $P_-$ is the projector onto the antisymmetric subspace of $\mathcal{H} \otimes \mathcal{H}$ of dimension $\frac{d(d-1)}{2}$. The covariance property of the channel (29) is

$$\tilde{\Phi}(U\rho U^*) = (U \otimes U)\tilde{\Phi}(\rho)(U^* \otimes U^*),$$

as follows from the fact that $P_-(U \otimes U) = (U \otimes U)P_-$.

## 5. A hierarchy of the additivity conjectures

**5.1. Convex closure.** To find out the intrinsic connection between the output entropy and the $\chi$-capacity, let us define the average $\bar{\rho}_\pi = \sum_x \pi(x)\rho(x)$ of the ensemble $\pi$ and rewrite the expression (5) in the form

$$C_\chi(\Phi) = \max_{\rho \in \mathfrak{S}(\mathcal{H})} \big[H(\Phi(\rho)) - \hat{H}_\Phi(\rho)\big], \tag{30}$$

where

$$\hat{H}_\Phi(\rho) = \min_{\pi:\bar{\rho}_\pi=\rho} \sum_x \pi(x) H(\Phi(\rho(x)))$$

is the *convex closure* [30] of the output entropy $H(\Phi(\rho))$[3]. The function $\hat{H}_\Phi(\rho)$ is a natural generalization of another important quantity in quantum information theory, namely the "entanglement of formation" [6] and reduces to it when the channel $\Phi$ is a partial trace. This quantity has the conjectured superadditivity property: *for an arbitrary state* $\rho_{12} \in \mathfrak{S}(\mathcal{H}_1 \otimes \mathcal{H}_2)$ *and arbitrary channels* $\Phi_1$, $\Phi_2$,

$$\hat{H}_{\Phi_1\otimes\Phi_2}(\rho_{12}) \geq \hat{H}_{\Phi_1}(\rho_1) + \hat{H}_{\Phi_2}(\rho_2), \tag{31}$$

*where* $\rho_1$, $\rho_2$ *are the partial traces of* $\rho_{12}$ *in* $\mathcal{H}_1$, $\mathcal{H}_2$ .

It is not difficult to see that this property implies additivity of both the minimal output entropy and the $\chi$-capacity:

**Proposition 5.1.** *The superadditivity property* (31) *implies the additivity properties* (12) *and* (6) *for given channels* $\Phi_1$, $\Phi_2$.

In the spirit of Theorem 4.1, one can prove ([25]) that

*if the relation* (31) *holds for the pair of CP maps* (*channels*) $\Phi_1$, $\Phi_2$, *then similar relation holds for the pair of their complementary maps* $\tilde{\Phi}_1$, $\tilde{\Phi}_2$. *If one of these relations holds for given* $\Phi_1$ *and arbitrary* $\Phi_2$, *then a similar relation holds for the complementary map* $\tilde{\Phi}_1$ *and arbitrary* $\Phi_2$.

Let $\{p_j\}$ be a finite probability distribution and let $\Phi_j : \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}'_j)$ be a collection of channels. The channel $\Phi\colon\ \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}\big(\sum_j \oplus \mathcal{H}'_j\big)$ is called *orthogonal convex sum* of the channels $\Phi_j$, if $\Phi(\rho) = \sum_j \oplus p_j \Phi_j(\rho)$ for all $\rho \in \mathfrak{S}(\mathcal{H})$.

**Proposition 5.2** ([26]). *Let* $\Phi_2$ *be an arbitrary channel. The properties* (11), (12), (31) *hold if* $\Phi_1$ *is an orthogonal convex sum of either an ideal channel or completely depolarizing channel and a channel* $\Phi^{(0)}$ *such that the corresponding property holds for* $\Phi^{(0)}$ *and* $\Phi_2$.

It follows that such a $\Phi_1$ fulfils the additivity of $\chi$-capacity (6).

In this way, for example, one obtains all the additivity properties for the important case of the *erasure channel*

$$\Phi(\rho) = \begin{bmatrix} p\rho & 0 \\ 0 & (1-p)\operatorname{Tr}\rho \end{bmatrix},$$

as it is the orthogonal convex sum of an ideal and a completely depolarizing channel.

[3] Here the same comment applies as to the attainability of the maximum in (5).

**5.2. Additivity for constrained channels.** In this section we consider several equivalent formulations of the additivity conjecture for channels with arbitrarily constrained inputs [26], which formally is substantially stronger than additivity of the unconstrained $\chi$-capacity. Let us denote

$$\chi_\Phi(\rho) = H(\Phi(\rho)) - \hat{H}_\Phi(\rho), \tag{32}$$

then the function $\chi_\Phi(\rho)$ is continuous and concave on the set $\mathfrak{S}(\mathcal{H})$ of all states in $\mathcal{H}$.

Consider the constraint on the ensemble $\pi$ with the average $\bar{\rho}_\pi = \sum_x \pi(x)\rho(x)$, defined by the requirement $\bar{\rho}_\pi \in \mathcal{A}$, where $\mathcal{A}$ is a closed subset of states. A particular case is the linear constraint $\mathcal{A} = \{\rho : \operatorname{Tr}\rho A \le \alpha\}$ for a positive operator $A$ and a number $\alpha \ge 0$. Define the *$\chi$-capacity of the $\mathcal{A}$-constrained channel* $\Phi$ by

$$C_\chi(\Phi; \mathcal{A}) = \max_{\rho\in\mathcal{A}} \chi_\Phi(\rho). \tag{33}$$

Note that the $\chi$-capacity for the unconstrained channel is $C_\chi(\Phi) = C(\Phi; \mathfrak{S}(\mathcal{H}))$. On the other hand, $\chi_\Phi(\rho) = C_\chi(\Phi; \{\rho\})$.

Let $\Phi_1$, $\Phi_2$ be two channels with the constraints $\mathcal{A}_1$, $\mathcal{A}_2$. For the channel $\Phi_1 \otimes \Phi_2$ we introduce the constraint $\mathcal{A}_1 \otimes \mathcal{A}_2 \equiv \{\rho : \operatorname{Tr}_{\mathcal{H}_2}\rho \in \mathcal{A}_1, \operatorname{Tr}_{\mathcal{H}_1}\rho \in \mathcal{A}_2\}$ and consider the conjecture

$$C_\chi(\Phi_1 \otimes \Phi_2; \mathcal{A}_1 \otimes \mathcal{A}_2) = C_\chi(\Phi_1; \mathcal{A}_1) + C_\chi(\Phi_2; \mathcal{A}_2), \tag{34}$$

which apparently implies (6).

**Theorem 5.3.** *Let $\Phi_1$ and $\Phi_2$ be two fixed channels. The following properties are equivalent:*

(i) *Equality* (34) *holds for arbitrary linear constraints $\mathcal{A}_1$, $\mathcal{A}_2$.*

(ii) *Equality* (34) *holds for arbitrary closed $\mathcal{A}_1$, $\mathcal{A}_2$.*

(iii) *For arbitrary $\rho_{12} \in \mathfrak{S}(\mathcal{H}_1 \otimes \mathcal{H}_2)$,*

$$\chi_{\Phi_1\otimes\Phi_2}(\rho_{12}) \le \chi_{\Phi_1}(\rho_1) + \chi_{\Phi_2}(\rho_2). \tag{35}$$

(iv) *Inequality* (31) *holds for arbitrary $\rho_{12} \in \mathfrak{S}(\mathcal{H}_1 \otimes \mathcal{H}_2)$.*

Here each property is easily seen to imply the preceding one, while the implication (i) $\Rightarrow$ (iv) is nontrivial. By Proposition 5.1 any of these properties imply the additivity properties (12), (6).

**5.3. The convex duality formulation.** In [4], tools from convex analysis were applied to study the relation of the additivity problem to superadditivity of entanglement

of formation. Here we apply a similar approach to the conjecture (31). Given a channel $\Phi$, its output entropy $H(\Phi(\rho))$ is a continuous concave function on the state space $\mathfrak{S}(\mathcal{H})$. Consider its modified Legendre transform

$$\begin{aligned} H_\Phi^*(X) &= \min_{\rho\in\mathfrak{S}(\mathcal{H})}\{\operatorname{Tr}\rho X + H(\Phi(\rho))\} \\ &= \min_{\rho\in\mathfrak{S}(\mathcal{H})}\{\operatorname{Tr}\rho X + \hat{H}_\Phi(\rho)\}, \quad X\in\mathfrak{M}_h(\mathcal{H}), \end{aligned} \tag{36}$$

where $\mathfrak{M}_h(\mathcal{H})$ is a real normed space of Hermitian operators in $\mathcal{H}$.

Now let $\Phi_1$, $\Phi_2$ be two channels.

**Lemma 5.4.** *The superadditivity* (31) *of the convex closure $\hat{H}_\Phi(\rho)$ is equivalent to the following additivity property of $H_\Phi^*(X)$:*

$$H^*_{\Phi_1\otimes\Phi_2}(X_1\otimes I_2 + I_1\otimes X_2) = H^*_{\Phi_1}(X_1) + H^*_{\Phi_2}(X_2), \tag{37}$$

*for all $X_1\in\mathfrak{M}_h(\mathcal{H}_1)$, $X_2\in\mathfrak{M}_h(\mathcal{H}_2)$.*

Since $H_\Phi^*(0) = \check{H}(\Phi)$, by letting $X_1 = X_2 = 0$, the relation (37) implies additivity of the minimal output entropy (12).

**5.4. The global equivalence.** A remarkable result was obtained by Shor [48] who showed that different forms of the additivity conjecture become equivalent if one considers their validity for *all* channels. Here we describe a basic construction from [48] which in combination with Proposition 5.1 and Theorem 5.3 suffices for the proof of the following result.

**Theorem 5.5.** *The conjectures* (6), (12), (31), (34) *are globally equivalent in the sense that if one of them holds true for all channels $\Phi_1$, $\Phi_2$, then any of the others is also true for all channels.*

Let us argue that if additivity of the minimal output entropy (12) holds for all channels, then (37) holds for all channels. By Lemma 5.4 this will imply (31) and hence, by Theorem 5.3, all the other properties.

First of all we observe that $H_\Phi^*(X+\lambda I) = H_\Phi^*(X) + \lambda$, which implies that it is sufficient to establish (37) only for $X_1, X_2 \geq 0$. The idea of proof is to build, for any channel $\Phi$ and $X\geq 0$, a sequence of channels $\Phi'_{X,n}$ such that

$$\min_\rho H(\Phi'_{X,n}(\rho)) \equiv \check{H}(\Phi'_{X,n}) = \min_\rho[H(\Phi(\rho)) + \operatorname{Tr}\rho X] + o(1) \equiv H_\Phi^*(X) + o(1).$$

One can then apply the convex duality argument to deduce for the original channels the additivity property (37), which is equivalent to (31), from the additivity of the minimal output entropy for channels $\Phi'_{X,n}$.

Given a channel $\Phi\colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}')$ and a positive $X\in\mathcal{H}$, the new channel $\Phi'_{X,n}$ is constructed as follows. Choose a constant $c\geq\|X\|$, then $E = c^{-1}X$ satisfies $0\leq E\leq I$. Let $q_n\in(0,1)$ be such that

$$(1-q_n)\log n = c, \quad n = 2, 3, \ldots.$$

Then $\Phi'_{X,n}\colon \mathfrak{M}(\mathcal{H}) \to \mathfrak{M}(\mathcal{H}'_n)$, where $\mathcal{H}'_n = \mathcal{H}' \oplus \mathcal{H}_n \oplus \mathbb{C}$, acts on $\rho \in \mathfrak{M}(\mathcal{H})$ as follows:

$$\Phi'_{X,n}(\rho) = \begin{bmatrix} q_n\Phi(\rho) & 0 & 0 \\ 0 & (1-q_n)(\operatorname{Tr}\rho E)\frac{I_n}{n} & 0 \\ 0 & 0 & (1-q_n)\operatorname{Tr}\rho(I-E) \end{bmatrix}. \tag{38}$$

This is an orthogonal convex sum of CP maps, preserving trace, and hence is a channel.

The intuition is that the action of $\Phi'_{X,n}(\rho)$ can be described as follows. With probability $q_n$ (which tends to 1 as $n \to \infty$) it acts as the channel $\Phi$, resulting in the state $\Phi(\rho)$. With probability $(1 - q_n)$, however, a quantum measurement described by the resolution of the identity (the quantum observable) $\{E, I - E\}$ is made, so that the first outcome appears with probability $\operatorname{Tr}\rho E$, while the second appears with probability $\operatorname{Tr}\rho(I - E)$. In the first case the output is the chaotic state $\frac{I_n}{n}$ in the $n$-dimensional unitary space $\mathcal{H}_n$; in the second case the output is a pure state orthogonal to $\mathcal{H}' \oplus \mathcal{H}_n$. In this way the channel $\Phi'_{X,n}(\rho)$ with high probability $q_n$ acts as the initial channel, while with a small probability $(1 - q_n)(\operatorname{Tr}\rho E)$ outputs a high dimensional chaotic state $\frac{I_n}{n}$, providing the knowledge about the value of $\operatorname{Tr}\rho E = c^{-1}\operatorname{Tr}\rho X$ involved in the definition of $H^*_\Phi(X)$. This is formalized by proving the uniform estimate

$$H\big((\Phi'_{X,n} \otimes \Phi_2)(\rho_{12})\big) = q_n H\big((\Phi \otimes \Phi_2)(\rho_{12})\big) + \operatorname{Tr}\rho_1 X + o(1),$$

double application of which reduces the property (37) for initial channels to the additivity of the minimal output entropy for the channels $\Phi'_{X,n}$.

A modification of this construction can be also used to show that *if the unconstrained additivity* (6) *holds for all channels, then additivity* (34) *for all channels with arbitrary constraints holds as well* [26]. This completes the global equivalence.

## 6. Infinite-dimensional channels

We have seen that the additivity problem is not completely solved even for the minimal dimension 2. Nevertheless there are several good reasons to consider the problem in *infinite* dimensions.

There is an important and interesting class of Bosonic Gaussian channels, see [28], which act in infinite dimensional Hilbert space. Analysis of continuity properties of the entropic characteristics of an infinite-dimensional channel becomes important since, as is well known, the entropy may then have a rather pathological behavior. It is only lower semicontinuous and "almost everywhere" infinite in the infinite-dimensional case [52]. Another issue is the study of conditions for compactness of subsets of quantum states and ensembles, giving a key for attainability of extrema in expressions for the capacity and the convex closure of the output entropy.

The proof of global equivalence of different forms of the additivity conjecture for finite dimensional channels (Section 5.4), using infinitely growing channel extensions

in fact relies upon the discontinuity of the $\chi$-capacity as a function of the channel in infinite dimensions. This also calls for a study of continuity properties of the entropic quantities related to the classical capacity of infinite dimensional channels. Such a study was undertaken in a series of works [23], [27], [46]. In particular it was shown that in spite of the aforementioned discontinuities, additivity for all finite-dimensional channels implies additivity of the $\chi$-capacity of infinite-dimensional channels with arbitrary constraints [46].

There are two important features essential for channels in infinite dimensions. One is the necessity of the input constraints (such as mean energy constraint for Gaussian channels) to prevent from infinite capacities (although considering input constraints was shown quite useful also in the study of the additivity conjecture for channels in finite dimensions [26]). The other is the natural appearance of infinite, and, in general, "continuous" state ensembles understood as probability measures on the set of all quantum states. By using compactness criteria from probability and operator theory one can show that the set of all such generalized ensembles with the barycenter in a compact set of states is itself weakly compact. With this in hand a sufficient condition for existence of an optimal generalized ensemble for a constrained quantum channel can be given. This condition can be efficiently verified in the case of Bosonic Gaussian channels with constrained mean energy [27].

However apart from mere existence one would like to have an explicit description of the optimal states and ensembles in the case of quantum Gaussian channels. In classical information theory Gaussian channels have Gaussian maximizers, and there is an analytical counterpart of this phenomenon for $(q \to p)$-norms of integral operators with Gaussian kernels, see [39]. Whether a similar description holds true for Bosonic Gaussian channels is another open question (for some partial results in this direction see [28], [17], [18], [55]). We only mention here that a positive solution of this question may also depend on the validity of the multiplicativity/ additivity conjecture [39], [55].

## References


[1] Alicki, R., Fannes, M., Note on multiple additivity of minimal Rényi entropy output of the Werner-Holevo channels. quant-ph/0407033.

[2] Amosov, G. G., Holevo, A. S., On the multiplicativity conjecture for quantum channels. *Theor. Probab. Appl.* **47** (1) (2002), 143–146.

[3] Amosov, G. G., Holevo, A. S., and Werner, R. F., On some additivity problems in quantum information theory. *Probl. Inform. Transm.* **36** (4) (2000), 25–34.

[4] Audenaert, K. M. R., Braunstein, S. L., On strong superadditivity of the entanglement of formation. *Commun. Math. Phys.* **246** (2004), 443–452.

[5] Bennett, C. H., Fuchs, C. A., Smolin, J. A., Entanglement-enhanced classical communication on a noisy quantum channel. In *Quantum Communication, Computing and Measurement, Proc. QCM96* (ed. by O. Hirota, A. S. Holevo and C. M. Caves), Plenum, New York 1997, 79–88.

[6] Bennett, C. H., Shor, P. W., Quantum information theory. *IEEE Trans. Inform. Theory* **44** (1998), 2724–2742.

[7] Bhatia, R., *Matrix Analysis*. Grad. Texts in Math. 169, Springer-Verlag, New York 1997.

[8] Carlen, E. A., Lieb, E. H., A Minkowski type trace inequality and strong subadditivity of quantum entropy. In *Differential operators and spectral theory* (ed. by V. Buslaev, M. Solomyak and D. Yafaev), Amer. Math. Soc. Transl. Ser. 2 189, Amer. Math. Soc., Providence, RI, 1999, 59–68.

[9] Choi, M.-D., Completely positive maps on complex matrices. *Linear Algebra Appl.* **10** (1975), 285–290.

[10] Cover, T. M., Thomas, J. A., *Elements of Information Theory*. J. Wiley and Sons, New York 1991.

[11] Datta, N., Holevo, A. S., Complementarity and additivity for depolarizing channels. quant-ph/0510145.

[12] Datta, N., Holevo, A. S., Suhov, Y. M., A quantum channel with additive minimum output entropy. quant-ph/0408176.

[13] Datta, N., Multiplicativity of maximal $p$-norms in Werner-Holevo channels for $1 \leq p \leq 2$. quant-ph/0410063.

[14] Devetak, I., Hayden, A. W., Winter, A., A resource framework for quantum Shannon theory. quant-ph/0512015.

[15] Devetak, I., Junge, M., King, C., Ruskai, M. B., Multiplicativity of completely bounded p-norms implies a new additivity result. quant-ph/0506196.

[16] Fukuda, M., Holevo, A. S., On Weyl-covariant channels. quant-ph/0510148.

[17] Giovannetti, V., Lloyd, S., Maccone, L., Shapiro, J. H., Yen, B. J., Minimum Rényi and Wehrl entropies at the output of bosonic channels. quant-ph/0404037.

[18] Giovannetti, V., Lloyd, S., Additivity properties of a Gaussian channel. quant-ph/0403075.

[19] Holevo, A. S., The capacity of quantum communication channel with general signal states. *IEEE Trans. Inform. Theory* **44** (1) (1998), 269–272.

[20] Holevo, A. S., Quantum coding theorems. *Russ. Math. Surveys* **53** (1998), 1295–1331.

[21] Holevo, A. S., *Statistical structure of quantum theory*. Lect. Notes Phys. Monogr. 67, Springer-Verlag, Berlin 2001.

[22] Holevo, A. S., *An introduction to quantum information theory*. MCCME (Moscow Independent University), Moscow 2002.

[23] Holevo, A. S., Entanglement-assisted capacities of constrained quantum channels. *Theory Probab. Appl.* **48** (2003), 243–255.

[24] Holevo, A. S., Additivity conjecture and covariant channels. *Int. J. Quant. Inform.* **3** (1) (2005), 41–48.

[25] Holevo, A. S., On complementary channels and the additivity problem. quant-ph/0509101.

[26] Holevo, A. S., Shirokov, M. E., On Shor's channel extension and constrained channels. *Commun. Math. Phys.* **249** (2004), 417–430.

[27] Holevo, A. S., Shirokov, M. E., Continuous ensembles and the $\chi$-capacity of infinite-dimensional channels. quant-ph/0403072.

[28] Holevo, A. S., Werner, R. F., Evaluating capacities of Bosonic Gaussian channels. *Phys. Rev. A.* **63** (2001), 032312.

[29] Horodecki, M., Shor, P. W., Ruskai, M. B., General entanglement breaking channels. *Rev. Math. Phys.* **15** (2003), 629–641.

[30] Magaril-Il'yaev, G. G., Tikhomirov, B. M., *Convex analysis: theory and applications*. Editorial URSS, Moscow 2000, Transl. Math. Monogr. 222, Amer. Math. Soc, Providence, RI, 2003.

[31] King, C., An application of the matrix inequality in quantum information theory. quant-ph/0412046.

[32] King, C., Maximal p-norms of entanglement breaking channels. quant-ph/0212057.

[33] King, C., Additivity for a class of unital qubit channels. quant-ph/0103156.

[34] King, C., The capacity of the quantum depolarizing channel. quant-ph/0204172.

[35] King, C., Koldan, N., New multiplicativity results for qubit maps. quant-ph/0512185.

[36] King, C., Matsumoto, K., Natanson, M., and Ruskai, M. B., Properties of conjugate channels with applications to additivity and multiplicativity. quant-ph/0509126.

[37] King, C., Nathanson, M., Ruskai, M.-B., Multiplicativity properties of entrywise positive maps on matrix algebras. quant-ph/0409181.

[38] King, C., Ruskai, M.-B., Comments on multiplicativity of maximal $p$-norms when $p = 2$. *Quantum Inf. Comput.* **4** (2004), 500–512.

[39] Lieb, E. H., Gaussian kernels have only Gaussian maximizers, *Invent. Math.* **102** (1990), 179–208.

[40] Lieb, E. H., Thirring, W. E., Inequalities for the moments of the eigenvalues of the Schrödinger Hamiltonian and their relation to Sobolev inequalities. In *Essays in Honor of Valentine Bargmann* (ed. by E. H. Lieb, B. Simon, A. Wightman), Stud. Math.Phys., Princeton University Press, 1976, 269–297.

[41] Matsumoto, K., Yura, F., Entanglement cost of antisymmetric states and additivity of capacity of some channels. quant-ph/0306009.

[42] Nielsen, M. A., Chuang, I., *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge 2000.

[43] Paulsen, V., *Completely bounded maps and operator algebras*. Cambridge University Press, Cambridge 2002.

[44] Pisier, G., *Non-Commutative Vector Valued $L_p$-spaces and Completely $p$-summing Maps*. *Astérisque* **247** (1998).

[45] Ruskai, M. B., Szarek, S., Werner, E., An analysis of completely-positive trace-preserving maps on $2 \times 2$ matrices. *Linear Algebra Appl.* **347** (2002), 159–187.

[46] Shirokov, M. E., The Holevo capacity of infinite dimensional channels. *Commun. Math. Phys.* **262** (2006), 137–159.

[47] Shor, P. W., Additivity of the classical capacity of entanglement-breaking quantum channels. *J. Math. Phys.* **43** (2003), 4334–4340.

[48] Shor, P. W., Equivalence of additivity questions in quantum information theory. *Commun. Math. Phys.* **246** (2004), 453–472.

[49] Schumacher, B., Westmoreland,M. D. , Sending classical information via noisy quantum channel. *Phys. Rev. A.* **56** (1) (1997), 131–138.

[50] Schumacher, B., Westmoreland, M. D. , Optimal signal ensembles. *Phys. Rev. A* **63** (2001), 022308.

[51] Stinespring, W. F., Positive functions on $C^*$-algebras. *Proc. Amer. Math. Soc.* **6** (1955), 211–316.

[52] Wehrl, A., General properties of entropy. *Rev. Mod. Phys.* **50** (1978), 221–250.

[53] Werner, R. F., Holevo, A. S., Counterexample to an additivity conjecture for output purity of quantum channels. *J. Math. Phys.* **43** (2002), 4353.

[54] Wolf, M. M., Eisert, J., Classical information capacity of a class of quantum channels. quant-ph/0412133.

[55] Wolf, M. M., Giedke, G., Cirac, J. I., Extremality of Gaussian quantum states. quant-ph/0509154.

Steklov Mathematical Institute, Gubkina 8, 119991 Moscow, Russian Federation

E-mail: holevo@mi.ras.ru

# Complex networks and decentralized search algorithms

Jon Kleinberg*

**Abstract.** The study of complex networks has emerged over the past several years as a theme spanning many disciplines, ranging from mathematics and computer science to the social and biological sciences. A significant amount of recent work in this area has focused on the development of random graph models that capture some of the qualitative properties observed in large-scale network data; such models have the potential to help us reason, at a general level, about the ways in which real-world networks are organized.

We survey one particular line of network research, concerned with small-world phenomena and decentralized search algorithms, that illustrates this style of analysis. We begin by describing a well-known experiment that provided the first empirical basis for the "six degrees of separation" phenomenon in social networks; we then discuss some probabilistic network models motivated by this work, illustrating how these models lead to novel algorithmic and graph-theoretic questions, and how they are supported by recent empirical studies of large social networks.

**Mathematics Subject Classification (2000).** Primary 68R10; Secondary 05C80, 91D30.

**Keywords.** Random graphs, complex networks, search algorithms, social network analysis.

## 1. Introduction

Over the past decade, the study of complex networks has emerged as a theme running through research in a wide range of areas. The growth of the Internet and the World Wide Web has led computer scientists to seek ways to manage the complexity of these networks, and to help users navigate their vast information content. Social scientists have been confronted by social network data on a scale previously unimagined: datasets on communication within organizations, on collaboration in professional communities, and on relationships in financial domains. Biologists have delved into the interactions that define the pathways of a cell's metabolism, discovering that the network structure of these interactions can provide insight into fundamental biological processes. The drive to understand all these issues has resulted in what some have called a "new science of networks" – a phenomenological study of networks as they arise in the physical world, in the virtual world, and in society.

At a mathematical level, much of this work has been rooted in the study of *random graphs* [14], an area at the intersection of combinatorics and discrete probability that

*Supported in part by a David and Lucile Packard Foundation Fellowship, a John D. and Catherine T. MacArthur Foundation Fellowship, and NSF grants CCF-0325453, IIS-0329064, CNS-0403340, and BCS-0537606.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

is concerned with the properties of graphs generated by random processes. While this has been an active topic of study since the work of Erdös and Rényi in the 1950s [26], the appearance of rich, large-scale network data in the 1990s stimulated a tremendous influx of researchers from many different communities. Much of this recent cross-disciplinary work has sought to develop random graph models that more tightly capture the qualitative properties found in large social, technological, and information networks; in many cases, these models are closely related to earlier work in the random graphs literature, but the issues arising in the motivating applications lead to new types of mathematical questions. For surveys covering different aspects of this general area, and in particular reflecting the various techniques of some of the different disciplines that have contributed to it, we refer the reader to recent review papers by Albert and Barabási [4], Bollobás [15], Kleinberg and Lawrence [39], Newman [52], and Strogatz [60], the volume of articles edited by Ben-Naim et al. [10], and the monographs by Dorogovtsev and Mendes [23] and Durrett [25], as well as books by Barabási [8] and Watts [62] aimed at more general audiences.

What does one hope to achieve from a probabilistic model of a complex network arising in the natural or social world? A basic strategy pursued in much of this research is to define a stylized network model, produced by a random mechanism that reflects the processes shaping the real network, and to show that this stylized model reproduces properties observed in the real network. Clearly the full range of factors that contribute to the observed structure will be too intricate to be fully captured by any simple model. But a finding based on a random graph formulation can help argue that the observed properties may have a simple underlying basis, even if their specifics are very complex. While it is crucial to realize the limitations of this type of activity – and not to read too much into the detailed conclusions drawn from a simple model – the development of such models has been a valuable means of proposing concrete, mathematically precise hypotheses about network structure and evolution that can then serve as starting points for further empirical investigation. And at its most effective, this process of modeling via random graphs can suggest novel types of qualitative network features – structures that people had not thought to define previously, and which become patterns to look for in new network datasets.

In the remainder of the present paper, we survey one line of work, motivated by the "small-world phenomenon" and some related search problems, that illustrates this style of analysis. We begin with a striking experiment by the social psychologist Stanley Milgram that frames the empirical issues very clearly [50], [61]; we describe a sequence of models based on random graphs that capture aspects of this phenomenon [64], [36], [37], [38], [63]; and we then discuss recent work that has identified some of the qualitative aspects of these models in large-scale network data [1], [43], [49]. We conclude with some further extensions to these random graph models, discussing the results and questions that they lead to.

## 2. The small-world phenomenon

The small-world phenomenon – the principle that we are all linked by short chains of acquaintances, or "six degrees of separation" [29] – has long been the subject of anecdotal fascination among the general public, and more recently has become the subject of both experimental and theoretical research. At its most basic level, it is a statement about networks, and human social networks in particular; it concerns the graph with one node corresponding to each person in the world, and an edge joining two people if they know each other on a first-name basis. When we say that this graph is a "small world," we mean, informally, that almost every pair of nodes is connected by a path with an extremely small number of steps.

One could worry about whether this graph is precisely specified – for example, what exactly it means to know someone on a first-name basis – but however one fixes a working definition for this, it is clear that the resulting graph encodes an enormous amount of information about society in general. It is also clear that it would be essentially impossible to determine its structure precisely. How then could one hope to test, empirically, the claim that most pairs of nodes in this graph are connected by short paths?

The social psychologist Stanley Milgram [50], [61] took up this challenge in the 1960s, conducting an experiment to test the small-world property by having people explicitly construct paths through the social network defined by acquaintanceship. To this end, he chose a *target person* in the network, a stockbroker living in a suburb of Boston, and asked a collection of randomly chosen "starter" individuals each to forward a letter to the target. He provided the target's name, address, occupation, and some personal information, but stipulated that the participants could not mail the letter directly to the target; rather, each participant could only advance the letter by forwarding it to a single acquaintance that he or she knew on a first-name basis, with the goal of reaching the target as rapidly as possible. Each letter thus passed successively from one acquaintance to another, closing in on the stockbroker outside Boston.

The letters thus acted as virtual "tracers," mapping out paths through the social network. Milgram found that the median length among the completed paths was six, providing the first concrete evidence for the abundance of short paths connecting far-flung pairs of individuals in society, as well as supplying the basis for the number "six" in the resulting pop-cultural mantra. One needs to be careful in interpreting this finding, of course: many of the chains never reached the target, and the target himself was a relatively "high-status" individual who may have been easier to reach than an arbitrary person (see e.g. the recent critique by Kleinfeld [41]). But since Milgram's work, the overall conclusion has been accepted at least at a qualitative level: social networks tend to exhibit very short paths between essentially arbitrary pairs of nodes.

## 3. Basic models of small-world networks

Why should social networks exhibit this type of a small-world property? Earlier we suggested that interesting empirical findings about networks often motivate the development of new random graph models, but we have to be careful in framing the issue here: a simple abundance of short paths is in fact something that most basic models of random graphs already "get right." As a paradigmatic example of such a result, consider the following theorem of Bollobás and de la Vega [14], [17].

**Theorem 3.1** ([17]). *Fix a constant $k \geq 3$. If we choose uniformly at random from the set of all $n$-node graphs in which each node has degree exactly $k$, then with high probability every pair of nodes will be joined by a path of length $O(\log n)$.*

(Following standard notation and terminology, we say that the *degree* of a node is the number of edges incident to it. We say that a function is $O(f(n))$ if there is a constant $c$ so that for all sufficiently large $n$, the function is bounded by $cf(n)$.) In fact, [17] states a much more detailed result concerning the dependence on $n$, but this will not be crucial for our purposes here.

Path lengths that are logarithmic in $n$ – or more generally *polylogarithmic*, bounded by a polynomial function of $\log n$ – will be our "gold standard" in most of this discussion. We will keep the term "small world" itself informal; but we will consider a graph to be a small world, roughly, when all (or most) pairs of nodes are connected by paths of length polylogarithmic in $n$, since in such a case the path lengths are exponentially smaller than the number of nodes.

Watts and Strogatz [64] argued that there is something crucial missing from the picture provided by Theorem 3.1. A standard random graph (for example, as in Theorem 3.1) is locally very sparse; with reasonably high probability, none of the neighbors of a given node $v$ are themselves neighbors of one another. But this is far from true in most naturally occurring networks: in real network data, many of a node's neighbors are joined to each other by edges. (For example, in a social network, many of our friends know each other.) Indeed, at an implicit level, this is a large part of what makes the small-world phenomenon surprising to many people when they first hear it: the social network appears from the local perspective of any one node to be highly "clustered," rather than the kind of branching tree-like structure that would more obviously reach many nodes along very short paths.

Thus, Watts and Strogatz proposed thinking about small-world networks as a kind of superposition: a structured, high-diameter network with a relatively small number of "random" links added in. As a model for social networks, the structured underlying network represents the "typical" social links that we form with the people who live near us, or who work with us; the additional random links are the chance, long-range connections that play a large role in creating short paths through the network as a whole.

This kind of hybrid random graph model had been studied earlier by Bollobás and Chung [16]; they showed that a small density of random links can indeed produce

short paths very effectively. In particular they proved the following, among other results.

**Theorem 3.2** ([16]). *Consider a graph $G$ formed by adding a random matching to an $n$-node cycle.* (*In other words, we assume $n$ is even, pair up the nodes on the cycle uniformly at random, and add edges between each of these node pairs.*) *With high probability, every pair of nodes will be joined by a path of length $O(\log n)$.*

Here too, Bollobás and Chung in fact proved a much more detailed bound on the path lengths; see [16] for further details.

This is quite close to the setting of the Watts-Strogatz work, who also considered cycles with random matchings as a model system for analysis. For our purposes here, we will begin with the following *grid-based model*, which is qualitatively very similar. We start with a two-dimensional $n \times n$ grid graph, and then for each node $v$, we add one extra directed edge to some other node $w$ chosen uniformly at random. (We will refer to $w$ as the *long-range contact* of $v$; to distinguish this, we will refer to the other neighbors of $v$, defined by the edges of the grid, as its *local contacts*.) Following the Watts-Strogatz framework, one can interpret this model as a metaphor for a social network embedded in an underlying physical space – people tend to know their geographic neighbors, as well as having friendships that span long distances. It is also closely related to *long-range percolation models*, though the questions we consider are fairly different; we discuss these connections in Section 7. For the present discussion, though, the essential feature of this model is its superposition of structured and random links, and it is important to note that the results to follow carry over directly to a wide range of variations on the model. Indeed, a significant part of what follows will be focused on a search for the most general framework in which to formulate these results.

## 4. Decentralized search in small-world networks

Thus far we have been discussing purely structural issues; but if one thinks about it, the original Milgram experiment contains a striking algorithmic discovery as well: not only did short paths exist in the social network, but people, using knowledge only of their own acquaintances, were able to collectively construct paths to the target. This was a necessary consequence of the way Milgram formulated the task for his participants; if one really wanted the *shortest* path from a starting person to the target, one should have instructed the starter to forward a letter to *all* of his or her friends, who in turn should have forwarded the letter to all of their friends, and so forth. This "flooding" of the network would have reached the target as rapidly as possible; but for obvious reasons, such an experiment was not a feasible option. As a result, Milgram was forced to embark on the much more interesting experiment of constructing paths by "tunneling" through the network, with the letter advancing just one person at a

time – a process that could well have failed to reach the target, even if a short path existed.

This algorithmic aspect of the small-world phenomenon raises fundamental questions – why should the social network have been structured so as to make this type of decentralized routing so effective? Clearly the network contained some type of "gradient" that helped participants guide messages toward the target, and this is something that we can try to model; the goal would be to see whether decentralized routing can be proved to work in a simple random-graph model, and if so, to try extracting from this model some qualitative properties that distinguish networks in which this type of routing can succeed. It is worth noting that these issues reach far beyond the Milgram experiment or even social networks; routing with limited information is something that takes place in communication networks, in browsing behavior on the World Wide Web, in neurological networks, and in a number of other settings – so an understanding of the structural underpinnings of efficient decentralized routing is a question that spans all these domains.

To begin with, we need to be precise about what we mean by a decentralized algorithm. In the context of the grid-based model in the previous section, we will consider algorithms that seek to pass a message from a starting node $s$ to a target node $t$, by advancing the message along edges. In each step of this process, the current message-holder $v$ has knowledge of the underlying grid structure, the location of the target $t$ on the grid, and its own long-range contact. The crucial point is that it does not know the long-range contacts of any other nodes. (Optionally, we can choose to have $v$ know the path taken by the message thus far, but this will not be crucial in any of the results to follow.) Using this information, $v$ must choose one of its network neighbors $w$ to pass the message to; the process then continues from $w$. We will evaluate decentralized algorithms according to their *delivery time* – the expected number of steps required to reach the target, over a randomly generated set of long-range contacts, and randomly chosen starting and target nodes. Our goal will be to find algorithms with delivery times that are polylogarithmic in $n$.

It is interesting that while Watts and Strogatz proposed their model without the algorithmic aspect in mind, it is remarkably effective as a simple system in which to study the effectiveness of decentralized routing. Indeed, to be able to pose the question in a non-trivial way, one wants a network that is partially known to the algorithm and partially unknown – clearly in the Milgram experiment, as well as in other settings, individual nodes use knowledge not just of their own local connections, but also of certain global "reference frames" (comparable to the grid structure in our setting) in which the network is embedded. Furthermore, for the problem to be interesting, the "known" part of the network should be likely to contain no short path from the source to the target, but there should be a short path in the full network. The Watts-Strogatz model combines all these features in a minimal way, and thus allows us to consider how nodes can use what they know about the network structure to construct short paths.

Despite all this, the first result here is negative.

**Theorem 4.1** ([36], [37]). *The delivery time of any decentralized algorithm in the grid-based model is* $\Omega(n^{2/3})$.

(We say that a function is $\Omega(f(n))$ if there is a constant $c$ so that for infinitely many $n$, the function is at least $cf(n)$.)

This shows that there are simple models in which there can be an exponential separation between the lengths of paths and the delivery times of decentralized algorithms to find these paths. However, it is clearly not the end of the story; rather, it says that the random links in the Watts-Strogatz model are somehow too "unstructured" to support the kind of decentralized routing that one found in the Milgram experiment. It also raises the question of finding a simple extension of the model in which efficient decentralized routing becomes possible.

To extend the model, we introduce one additional parameter $\alpha \geq 0$ that controls the extent to which the long-range links are correlated with the geometry of the underlying grid. First, for two nodes $v$ and $w$, we define their *grid distance* $\rho(v, w)$ to be the number of edges in a shortest path between them on the grid. The idea behind the extended model is to have the long-range contacts favor nodes at smaller grid distance, where the bias is determined by $\alpha$. Specifically, we define the *grid-based model with exponent* $\alpha$ as follows. We start with a two-dimensional $n \times n$ grid graph, and then for each node $v$, we add one extra directed edge to some other long-range contact; we choose $w$ as the long-range contact for $v$ with probability proportional to $\rho(v, w)^{-\alpha}$. Note that $\alpha = 0$ corresponds to the original Watts-Strogatz model, while large values of $\alpha$ produce networks in which essentially no edges span long distances on the grid.

We now have a continuum of models that can be studied, parameterized by $\alpha$. When $\alpha$ is very small, the long-range links are "too random," and can't be used effectively by a decentralized algorithm; when $\alpha$ is large, the long-range links appear to be "not random enough," since they simply don't provide enough of the long-distance jumps that are needed to create a small world. Is there an optimal operating point for the network, where the distribution of long-range links is sufficiently balanced between these extremes to be of use to a decentralized routing algorithm?

In fact there is; as the following theorem shows, there is a unique value of $\alpha$ in the grid-based model for which a polylogarithmic delivery time is achievable.

**Theorem 4.2** ([36], [37]). (a) *For* $0 \leq \alpha < 2$*, the delivery time of any decentralized algorithm in the grid-based model is* $\Omega(n^{(2-\alpha)/3})$.

(b) *For* $\alpha = 2$*, there is a decentralized algorithm with delivery time* $O(\log^2 n)$.

(c) *For* $\alpha > 2$*, the delivery time of any decentralized algorithm in the grid-based model is* $\Omega(n^{(\alpha-2)/(\alpha-1)})$.

(We note that the lower bounds in (a) and (c) hold even if each node has an arbitrary constant number of long-range contacts, rather than just one.)

The decentralized algorithm achieving the bound in (b) is very simple: each node simply forwards the message to a neighbor – long-range or local – whose grid distance to the target is as small as possible. (In other words, each node uses its long-range

contact if this gets the message closer to the target on the grid; otherwise, it uses a local contact in the direction of the target.) The analysis of this algorithm proceeds by showing that, for a constant $\varepsilon > 0$, there is a probability of at least $\varepsilon / \log n$ in every step that the grid distance to the target will be halved. It is also worth noting that the proof can be directly adapted to a grid in any constant number of dimensions; an analogous trichotomy arises, with polylogarithmic delivery time achievable only when $\alpha$ is equal to the dimension.

At a more general level, the proof of Theorem 4.2(b) shows that the crucial property of exponent $\alpha = 2$ is the following: rather than producing long-range contacts that are uniformly distributed over the grid (as one gets from exponent $\alpha = 0$), it produces long-range contacts that are approximately uniformly distributed over "distance scales": the probability that the long-range contact of $v$ is at a grid distance between $2^{j-1}$ and $2^j$ away from $v$ is approximately the same for all values of $j$ from 1 to $\log n$.

From this property, one sees that there is a reasonable chance of halving the message's grid distance to the target, independent of how far away it currently is. The property also has an intuitively natural meaning in the context of the original Milgram experiment; subject to all the other simplifications made in the grid model, it says very roughly that decentralized routing can be effective when people have approximately the same density of acquaintances at many different levels of distance resolution. And finally, this approximate uniformity over distance scales is the type of qualitative property that we mentioned as a goal at the outset. It is something that we can search for in other models and in real network data – tasks that we undertake in the next two sections.

## 5. Decentralized search in other models

**Hierarchical models.** A natural variation on the model of the previous section is to suppose that the network is embedded in a hierarchy rather than a grid – in other words, that the nodes reside at the leaves of a complete $b$-ary tree, and the underlying "distance" between two nodes is based on the height of their lowest common ancestor in this tree.

There are a number of settings where such a model suggests itself. To begin with, follow-up work on the Milgram experiment found that most decisions made by participants on how to forward the letter were based on one of two kinds of cues: geographical and occupational [35]. And if a two-dimensional grid is natural as a simple abstraction for the role of geography, then a hierarchy is a reasonable, also simple, approximation of the way in which people categorize occupations. Another domain in which hierarchies arise naturally is in the relationships among Web pages: for example, a Web page about sequence analysis of the yeast genome could be classified as being about genetics, more generally about biology, and more generally still about science, while a Web page reviewing performances of Verdi's *Aida* could

be classified as being about opera, more generally about music, and more generally still about the arts.

A natural assumption is that the density of links is lower for node pairs that are more widely separated in the underlying hierarchy, and this forms the basis for the following *hierarchical model with exponent* $\beta$. We begin with a complete $b$-ary tree having $n$ leaves (and hence of height $h = \log_b n$). For two leaves $v$ and $w$, let us define their *tree distance* $h(v, w)$ to be the height of their lowest common ancestor in the underlying tree. We now define the following random directed graph $G$ on the set $V$ of leaves: for a value $k$ and for each node $v$ in $V$, we construct $k$ edges out of $v$, choosing $w$ as the endpoint of the $i^{\text{th}}$ edge independently with probability proportional to $b^{-\beta h(v,w)}$. (We will refer to $k$ as the *out-degree* of the model.)

Thus, $\beta$ works much like $\alpha$ did in the grid-based model; when $\beta = 0$, we get uniform random selection, while larger values of $\beta$ bias the selection more toward "nearby" nodes. Now, in this case, a decentralized search algorithm is given the locations of a starting node $s$ and a target node $t$ in the hierarchy, and it must construct a path from $s$ to $t$, knowing only the edges out of nodes that it explicitly visits. Note that in defining the performance metric for a decentralized search algorithm in this model, we face a problem that we didn't encounter in the grid-based model: the graph $G$ may not contain a path from $s$ to $t$. Thus, we say that a decentralized algorithm here has delivery time $f(n)$ if, on a randomly generated $n$-node network, and with $s$ and $t$ chosen uniformly at random, the algorithm produces a path of length $O(f(n))$ with probability at least $1 - \varepsilon(n)$, where $\varepsilon(\cdot)$ is a function going to 0 as $n$ increases.

We now have the following analogue of Theorem 4.2, establishing that there is a unique value of $\beta$ for which polylogarithmic delivery time can be achieved when the network has polylogarithmic out-degree. This is achieved at $\beta = 1$, when the probability that $v$ links to a node at tree distance $h$ is almost uniform over choices of $h$. Also by analogy with the grid-based model, it suffices to use the simple "greedy" algorithm that always seeks to reduce the tree distance to the target by as much as possible.

**Theorem 5.1** ([38]). (a) *In the hierarchical model with exponent* $\beta = 1$ *and out-degree* $k = c \log^2 n$*, for a sufficiently large constant c, there is a decentralized algorithm with polylogarithmic delivery time.*

(b) *For every* $\beta \neq 1$ *and every polylogarithmic function* $k(n)$*, there is no decentralized algorithm in the hierarchical model with exponent* $\beta$ *and out-degree* $k(n)$ *that achieves polylogarithmic delivery time.*

Watts, Dodds, and Newman [63] independently proposed a model in which each node resides in several distinct hierarchies, reflecting the notion that participants in the small-world experiment were simultaneously taking into account several different notions of "proximity" to the target. Concretely, their model constructs a random graph $G$ as follows. We begin with $q$ distinct complete $b$-ary trees, for a constant $q$, and in each of these trees, we independently choose a random one-to-one mapping of the nodes onto the leaves. We then apply a version of the hierarchical model

above, separately in each of the trees; the result is that each node of $G$ acquires edges independently through its participation in each tree. (There are a few minor differences between their procedure within each hierarchy and the hierarchical model described above; in particular, they map multiple nodes to the same leaf in each hierarchy, and they generate each edge by choosing the tail uniformly at random, and then the head according to the hierarchical model. The result is that nodes will not in general all have the same out-degree.)

Precisely characterizing the power of decentralized search in this model, at an analytical level, is an open question, but Watts et al. describe a number of interesting findings obtained through simulation [63]. They study what is perhaps the most natural search algorithm, in which the current message-holder forwards the message to its neighbor who is closest (in the sense of tree distance) to the target in any of the hierarchies. Using an empirical definition of efficiency on networks of several hundred thousand nodes, they examined the set of $(\beta, q)$ pairs for which the search algorithm was efficient; they found that this "searchable region" was centered around values of $\beta \geq 1$ (but relatively close to 1), and on small constant values of $q$. (Setting $q$ equal to 2 or 3 yielded the widest range of $\beta$ for which efficient search was possible.) The resulting claim, at a qualitative level, is that efficient search is facilitated by having a small number of different ways to measure proximity of nodes, and by having a small bias toward nearby nodes in the construction of random edges.

**Models based on set systems.** One can imagine many other ways to construct networks in this general style – for example, placing nodes on both a hierarchy and a lattice simultaneously – and so it becomes natural to consider more general frameworks in which a range of these bounds on searchability might follow simultaneously from a single result. One such approach is based on constructing a random graph from an underlying set system, following the intuition that individuals in a social network often form connections because they are both members of the same small group [38]. In other words, two people might be more likely to form a link because they live in the same town, work in the same profession, have the same religious affiliation, or follow the work of the same obscure novelist.

Concretely, we start with a set of nodes $V$, and a collection of subsets $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ of $V$, which we will call the set of *groups*. It is hard to say much of interest for arbitrary set systems, but we would like our framework to include at least the collection of balls or subsquares in a grid, and the collection of rooted sub-trees in a hierarchy. Thus we consider set systems that satisfy some simple combinatorial properties shared by these two types of collections. Specifically, for constants $\lambda < 1$ and $\kappa > 1$, we impose the following three properties.

(i) The full set $V$ is one of the groups.

(ii) If $S_i$ is a group of size $g \geq 2$ containing a node $v$, then there is a group $S_j \subseteq S_i$ containing $v$ that is strictly smaller than $S_i$, but has size at least $\min(\lambda g, g - 1)$.

(iii) If $S_{i_1}, S_{i_2}, S_{i_3}, \ldots$ are groups that all have size at most $g$ and all contain a common node $v$, then their union has size at most $\kappa g$.

The most interesting property here is (iii), which can viewed as a type of "bounded growth" requirement; one can easily verify that it (along with (i) and (ii)) holds for the set of balls in a grid and the set of rooted sub-trees in a hierarchy.

Given a collection of groups, we construct a random graph as follows. For nodes $v$ and $w$, we define $g(v, w)$ to be the size of the smallest group containing both of them – this will serve as a notion of "distance" between $v$ and $w$. For a fixed exponent $\gamma$ and out-degree value $k$, we construct $k$ edges out of each node $v$, choosing $w$ as the endpoint of the $i^{\text{th}}$ edge from $v$ independently with probability proportional to $g(v, w)^{-\gamma}$. We will refer to this as the *group-based model* with set system $\mathcal{S}$, exponent $\gamma$, and out-degree $k$. A decentralized search algorithm in such a random graph is given knowledge of the full set system, and the identity of a target node; but it only learns the links out of a node $v$ when it reaches $v$. We now have the following theorem.

**Theorem 5.2** ([38]). (a) *Given an arbitrary set system $\mathcal{S}$ satisfying properties* (i), (ii), *and* (iii), *there is a decentralized algorithm with polylogarithmic delivery time in the group-based model with set system $\mathcal{S}$, exponent $\gamma = 1$, and out-degree $k = c \log^2 n$, for a sufficiently large constant $c$.*

(b) *For every set system $\mathcal{S}$ satisfying properties* (i), (ii), *and* (iii), *every $\gamma < 1$, and every polylogarithmic function $k(n)$, there is no decentralized algorithm achieving polylogarithmic delivery time in the group-based model with set system $\mathcal{S}$, exponent $\gamma$ and out-degree $k(n)$.*

In other words, efficient decentralized search is possible when nodes link to each other with probability inversely proportional to the size of the smallest group containing both of them. As a simple concrete example, if the groups are the balls in a two-dimensional grid, then the size of the smallest group containing two nodes at distance $\rho$ is proportional to $\rho^2$, and so the link probability indicated by Theorem 5.2 (a) is proportional to $\rho^{-2}$; this yields an analogue of Theorem 4.2 (b), the inverse-square result for grids. (The present setting is not exactly the same as the one there; here, we do not automatically include the edges of the original grid when constructing the graph, but we construct a larger number of edges out of each node.)

Simple examples show that one cannot directly formulate a general negative result in this model for the case of exponents $\gamma > 1$ [38]. At a higher level, the group-based model is clearly not the only way to generalize the results thus far; in the next section we will discuss one other recent approach, and the development of other general models is a natural direction for further research.

## 6. Design principles and network data

In addition to their formulation as basic questions about search algorithms in graphs, the models we have been discussing thus far have been used as design principles in

file-sharing systems; and they have been found to capture some of the large-scale structure of human social networks as reflected in on-line data.

**Peer-to-peer systems and focused web crawling.** A recurring theme in recent work on complex networks is the way in which simple probabilistic models can rapidly become design principles for new types of networked systems. In the case of small-world networks, one observes this phenomenon in the development of protocols for peer-to-peer file sharing. The design of such protocols has become an active topic of research in the area of computer systems, motivated in part by the explosion of popular interest in peer-to-peer applications following the emergence of Napster and music file-sharing in 1999. The goal of such applications was to allow a large collection of users to share the content residing on their personal computers, and in their initial conception, the systems supporting these applications were based on a centralized index that simply stored, in a single place, the files that all users possessed. This way, queries for a particular piece of content could be checked against this index, and routed to the computer containing the appropriate file.

The music-sharing application of these systems, of course, ran into significant legal difficulties; but independent of the economic and intellectual property issues raised by this particular application, it is clear that systems allowing large user communities to share content have a much broader range of potential, less controversial uses, provided they can be structured in a robust and efficient way. This has stimulated much subsequent study in the research community, focusing on *decentralized* approaches in which one seeks file-sharing solutions that do not rely on a single centralized index of all the content.

In this decentralized version of the problem, the crux of the challenge is clear: each user has certain files on his or her own computer, but there is no single place that contains a global list of all these files; if someone poses a query looking for a specific piece of content, how can we efficiently determine which user (if any) possesses a copy of it? Without a central index, we are in a setting very much like that of the Milgram experiment: users must pose the query to a subset of their immediate network neighbors, who in turn can forward the query to some of their neighbors, and so forth. And this is where small-world models have played a role: a number of approaches to this problem have tried to explicitly set up the network on which the protocol operates so that its structure makes efficient decentralized search possible. We refer the reader to the surveys by Aspnes and Shah [6] and Lua et al. [44] for general reviews of this body of work, and the work of Clarke et al. (as described in [32]), Zhang et al. [67], Malkhi et al. [45], and Manku et al. [46] for more specific discussions of the relationship to small-world networks.

A related set of issues comes up in the design of *focused Web crawlers*. Whereas standard Web search engines first compile an enormous index of Web pages, and then answer queries by referring to this index, a focused crawler attempts to locate pages on a specific topic by following hyperlinks from one page to another, without first compiling an index. Again, the underlying issue here is the design of decentralized

search algorithms, in this case for the setting of the Web: when searching for relevant pages without global knowledge of the network, what are the most effective rules for deciding which links to follow? Motivated by these issues, Menczer [49] studied the extent to which the hierarchical model described in the previous section captures the patterns of linkage in large-scale Web data, using the hierarchical organization of topics provided by the Open Directory.

**Social network data.** The previous two applications – peer-to-peer systems and focused Web crawling – are both concerned with the structure of computer and information networks, although in both cases there are obvious social forces underlying their construction. Recent work has also investigated the extent to which the models described in the previous sections are actually reflected in data on human social networks. In other words, these small-world models make very concrete claims about the ways in which networks should be organized to support efficient search, but it is not *a priori* clear whether or not naturally occurring networks are organized in such ways. Two recent studies of this flavor have both focused on social networks that exist in on-line environments – as with the previous applications, we again see an intertwining of social and technological networks, but in these cases the emphasis is on the social component, with the on-line aspect mainly providing an opportune means of performing fine-grained analysis on a large scale.

In one study of this flavor, Adamic and Adar [1] considered the e-mail network of a corporate research lab: they collected data over a period of time, and defined an edge between any two people who exchanged at least a certain number of messages during this period. They overlaid the resulting network on a set system representing the organizational structure, with a set for each subgroup of the lab's organizational hierarchy. Among other findings, they showed that the probability of a link between individuals $v$ and $w$ scaled approximately proportional to $g(v, w)^{-3/4}$, compared with the value $g(v, w)^{-1}$ for efficient search from Theorem 5.2(a). (As above, $g(v, w)$ denotes the size of the smallest group containing both $v$ and $w$.) Thus, interactions in their data spanned large groups at a slightly higher frequency than the optimum for decentralized search. Of course, the e-mail network was not explicitly designed to support decentralized search, although one can speculate about whether there were implicit factors shaping the network into a structure that was easy to search; in any case, it is interesting that the behavior of the links with respect to the collection of groups is approximately aligned with the form predicted by the earlier theorems.

An even closer correlation with the structure predicted for efficient search was found in a large-scale study by Liben-Nowell et al. [43]. They considered LiveJournal, a highly active on-line community with several million participants, in which members communicate with one another, update personal on-line diaries, and post messages to community discussions. LiveJournal is a particularly appealing domain for studying the geographic distribution of links, because members provide explicit links to their friends in the system, and a large subset (roughly half a million at the time of the study in [43]) also provide a hometown in the continental U.S. As a result, one has

the opportunity to investigate, over a very large population, how the density of social network links decays with distance.

A non-trivial technical challenge that must be overcome in order to relate this data to the earlier models is that the population density of the U.S. is extremely non-uniform, and this makes it difficult to interpret predictions based on a model in which nodes are distributed uniformly over a grid. The generalization to group structures in the previous section is one way to handle non-uniformity; Liben-Nowell et al. propose an alternative generalization, *rank-based friendships*, that they argue may be more suitable to the geographic data here [43]. In the rank-based friendship model, one has a set of $n$ people assigned to locations on a two-dimensional grid, where each grid node may have an arbitrary positive number of people assigned to it. By analogy with the grid-based model from Section 4, each person $v$ chooses a *local contact* arbitrarily in each of the four neighboring grid nodes, and then chooses an additional *long-range contact* as follows. First, $v$ ranks all other people in order of their distance to herself (breaking ties in some canonical way); we let $\text{rank}_v(w)$ denote the position of $w$ in $v$'s ordered list, and say that $w$ is at *rank* $r$ with respect to $v$. $v$ then chooses $w$ as her long-range contact with probability proportional to $1/\text{rank}_v(w)$.

Note that this model generalizes the grid-based model of Section 4, in the sense that the grid-based model with the inverse-square distribution corresponds to rank-based friendship in which there is one person resident at each grid node. However, the rank-based friendship construction is well-defined for any population density, and Liben-Nowell et al. prove that it supports efficient decentralized search in general. They analyze a decentralized greedy algorithm that always forwards the message to a grid node as close as possible to the target's; and they define the *delivery time* in this case to be the expected number of steps needed to reach the grid node containing the target. (So we can imagine that the task here is to route the message to the hometown of the target, rather than the target himself; this is also consistent with the data available from LiveJournal, which only localizes people to the level of towns.)

**Theorem 6.1** ([43]). *For an arbitrary population density on a grid, the expected delivery time of the decentralized greedy algorithm in the rank-based friendship model is* $O(\log^3 n)$.

On the LiveJournal data, Liben-Nowell et al. examine the fraction of friendships $(v, w)$ where $w$ is at rank $r$ with respect to $v$. They find that this fraction is very close to inverse linear in $r$, in close alignment with the predictions of the rank-based friendship model.

This finding is notable for several reasons. First, as with the e-mail network considered by Adamic and Adar, there is no *a priori* reason to believe that a large, apparently amorphous social network should correspond so closely to a distribution predicted by a simple model for efficient decentralized search. Second, geography is playing a strong role here despite the fact that LiveJournal is an on-line system in which there are no explicit limitations on forming links with people arbitrarily far away; as a result, one might have (incorrectly) conjectured that it would be difficult

to detect the traces of geographic proximity in such data. And more generally, the analytical results of this section and the previous ones have been based on highly stylized models that nonetheless make very specific predictions about the theoretical "optimum" for search; to see these concrete predictions approximately borne out on real social network data is striking, and it suggests that there may be deeper phenomena yet to be discovered here.

## 7. Further results on small-world networks and decentralized search

**Long-range percolation.** The grid-based models we have been considering are closely related to the problem of *long-range percolation*. In the basic version of long-range percolation, one takes the infinite $d$-dimensional integer lattice $\mathbb{Z}^d$, and for each pair of nodes $(v, w)$ one includes an undirected edge between them independently with probability $\rho(v, w)^{-\alpha}$, where $\rho(v, w)$ is the grid distance between $v$ and $w$ and $\alpha \geq 0$ is a parameter of the model. Note that there are some small differences from the grid-based model described in Section 4: the graph is infinite, it is undirected, its nodes do not all have the same degree, and it does not automatically include edges between nearest neighbors on the lattice. In addition to these, a broader difference is in the nature of the questions investigated, with the initial work on long-range percolation focusing on the range of parameters for which an infinite connected component is likely to exist [3], [51], [57].

Motivated in part by the interest in small-world networks, work on long-range percolation began to investigate diameter issues – the maximum $D$ for which every node is connected by a path of at most $D$ steps. Benjamini and Berger [11] studied this problem in one dimension, modifying the model so that the graph is finite (restricted to the integers $\{1, 2, \ldots, n\}$), and so that edges are guaranteed to exist between adjacent integers. (They also studied the case in which the distance $\rho(\cdot, \cdot)$ is defined by assuming that the integers are "wrapped" into a cycle, so that $\rho(i, j)$ is not $|j - i|$ but $\min(|j - i|, n - |j - i|)$.) Their work was followed by results of Coppersmith et al. [20] and Biskup [13], who obtained sharper bounds in some cases and considered higher-dimensional lattices as well, in which the node set is $\{1, 2, \ldots, n\}^d$. As a result of this work, we know that the diameter of the graph changes qualitatively at the "critical values" $\alpha = d$ and $\alpha = 2d$. In particular, with high probability, the diameter is constant when $\alpha < d$ (due in essence to a result of [12]), is proportional to $\log n / \log \log n$ when $\alpha = d$ [20], is polylogarithmic in $n$ when $d < \alpha < 2d$ (with an essentially tight bound provided in [13]), and is lower-bounded by a polynomial in $n$ when $\alpha > 2d$ [11], [20]. The case $\alpha = 2d$ is largely open, and conjectured to have diameter polynomial in $n$ with high probability [11], [13]. It is also open whether the diameter for $\alpha > 2d$ is in fact linear in $n$; this has been proved for the one-dimensional case [11] and conjectured to hold for higher dimensions as well [11], [13], [20].

This pair of transitions at $\alpha = d$ and $\alpha = 2d$ was observed in a somewhat different setting by Kempe et al. [34], resolving a conjecture of Demers et al. [21] on the behavior of *gossip algorithms*. In this model, there are nodes located on the finite

$d$-dimensional lattice $\{1, 2, \ldots, n\}^d$, and in each time step each node $v$ picks a single other node and tells everything it currently knows to $w$; node $w$ is selected as the recipient of this information with probability proportional to $\rho(v, w)^{-\alpha}$. Information originating at one node thus spreads to other nodes, relayed in an epidemic fashion over time. Now, if a single node $v$ initially possesses a new piece of information at time 0, how long will it take before knowledge of this information has spread to a given node $w$? The main result of [34] is that the time required for this is polylogarithmic in $n$ for $\alpha \leq d$, is polylogarithmic in $\rho(v, w)$ but independent of $n$ for $d < \alpha < 2d$, and is polynomial in $\rho(v, w)$ for $\alpha > 2d$. Here too the case $\alpha = 2d$ is not well understood, which is interesting because this transitional value has particular importance in applications of gossip algorithms to distributed computing systems [54]. (See [34] for partial results concerning $\alpha = 2d$.)

For the specific grid-based model described in Section 4, Martel and Nguyen showed that with high probability the diameter is proportional to $\log n$ for $\alpha \leq d$, in the $d$-dimensional case [48]. They also identified transitions at $\alpha = d$ and $\alpha = 2d$ analogous to the case of long-range percolation [53]. In particular, their results show that while decentralized search can construct a path of length $O(\log^2 n)$ when $\alpha = d$, there in fact exist paths that are shorter by a logarithmic factor. (Note also the contrast with the corresponding results for the long-range percolation model when $\alpha \leq d$; in the grid-based model, the out-degree of each node is bounded by a constant, so a diameter proportional to $\log n$ is the smallest one could hope for; in the case of long-range percolation, on the other hand, the node degrees will be unbounded, allowing for smaller diameters.)

**Decentralized search with additional information.** A number of papers have studied the power of decentralized search algorithms that are provided with small amounts of additional information [28], [42], [47], [48], [66]. Whereas the model of decentralized algorithms in Section 4 charged unit cost to the algorithm for each node visited, the models in these subsequent papers make the following distinction: a node may "consult" a small number of nearby nodes, and then based on what it learns from this consultation, it chooses a node to forward the messages to. In bounding the number of steps taken by the algorithm, only the message-forwarding operations are counted, not the consultation.

In particular, Lebhar and Schabanel [42] consider an algorithm in which the node currently holding the message consults a set $S$ of up to $O(\log n)$ nodes within a small number of steps of it; after this, it forwards the message along a path to the node $w$ in $S$ that is closest to the target in grid distance. They show that, in total, the expected number of nodes consulted by this process is $O(\log^2 n)$ (as in the decentralized algorithm from Section 4), and that the actual path constructed to the target has only $O(\log n(\log\log n)^2)$ steps.

Manku, Naor, and Wieder [47] consider a simpler algorithm in the long-range percolation model on the $d$-dimensional lattice $\{1, 2, \ldots, n\}^d$ with $\alpha = d$. Note that nodes here will have unbounded degrees – proportional to $\log n$ in expectation, rather

than constant as in the grid-based model. Manku et al. analyze a *neighbor-of-neighbor* search algorithm in which the current message-holder $v$ consults each of its neighbors to learn the set $S$ of all of *their* neighbors; $v$ then forwards the message along the two-step path to the node in $S$ that lies closest to the target. They show that with high probability, this algorithm produces a path to the target of at most $O(\log n / \log\log n)$ steps, matching the bound of Coppersmith et al. [20] on the diameter of this network. Moreover, they show that the basic greedy algorithm, which simply forwards the message to the neighbor closest to the target, requires an expected number of steps proportional to $\log n$ to reach the target. Thus, one step of lookahead provides an asymptotic improvement in delivery time; and since one step of lookahead yields path lengths matching the diameter, additional lookahead does not offer any further asymptotic improvements.

Thus, the results of Manku et al. provide a rather sharp characterization of the power of lookahead in the long-range percolation model at the exponent $\alpha = d$ that allows for efficient decentralized search; determining a similarly precise delineation on the power of lookahead in the grid-based model (extending the aforementioned results of Lebhar and Schabanel) is an interesting open question.

**Small-world networks built on arbitrary underlying graphs.** The results in Section 5 describe various methods for constructing searchable networks based on underlying structures other than $d$-dimensional grids. In several recent papers, a number of further structures have been proposed as "scaffolds" for small-world networks [9], [27], [31], [53], [59].

In principle, one can consider adding long-range edges to any underlying graph $G$; Fraigniaud [27] asks whether any $G$ can be converted through such a process into a network that is efficiently searchable by a greedy algorithm. Specifically, suppose we choose a distribution over long-range contacts for each node of $G$, and we use this to generate a random graph $G'$ by adding a single long-range edge out of each node of $G$. We then consider the natural greedy algorithm for forwarding the message to a target $t$: the current message-holder passes the message to a neighbor that has the shortest path to the target as measured in $G$ (not in $G'$). Is it the case that for every graph $G$, there is a distribution over long-range contacts such that this algorithm has a delivery time that is polylogarithmic in $n$?

This question is open in general; note that the challenge in resolving it comes from the fact that a single choice of distribution per node must work (in expectation) over any possible destination, and that even if the graph $G'$ has nicely-structured short paths, the search algorithm is constrained to behave "greedily" in the original graph $G$. Fraigniaud answers the question in the affirmative for graphs of bounded tree-width as well as graphs in which there is no induced cycle of greater than a fixed length [27]; he also discusses some respects in which such underlying graphs are qualitatively consistent with observed properties of social networks. Duchon et al. answer the question in the affirmative for graphs satisfying a certain "bounded growth rate" property [24].

Slivkins [59] considers a different setting, in which nodes are embedded in an underlying metric space. He shows that if the metric is *doubling*, in the sense that every ball can be covered by a constant number of balls of half the radius (see e.g. [7], [30]), then there is a model such that each node generates a polylogarithmic number of long-range contacts from specified distributions, and a decentralized algorithm is then able to achieve a polylogarithmic delivery time. (Some of the logarithmic dependence here is on the *aspect ratio* of the metric – the ratio of the largest to the smallest distance – but it is possible to avoid this dependence in the bound on the delivery time. See [59] for further details on this issue.)

Finally, other work has studied search algorithms that exploit differences in node degrees. There are indications that people navigating social structures, in settings such as small-world experiments, take into account the fact that certain of their acquaintances simply know a large number of people [22]. Similarly, in peer-to-peer networks, it is also the case that certain nodes have an unusually large number of neighbors, and may thus be more useful in helping to forward queries. Adamic et al. [2] formalize these considerations by studying a random graph model in which high-degree nodes are relatively abundant, and decentralized search algorithms only have access to information about degrees of neighboring nodes, not to any embedding of the graph (spatial or otherwise). Through simulation, they find that for certain models, knowledge of degrees provides an improvement in search performance.

Simsek and Jensen [58] consider a model which combines spatial embedding with variable node degrees. Specifically, they study a variant of the grid-based model from Section 4 in which nodes have widely varying degrees, and a decentralized algorithm has access both to the locations of its neighbors and to their degrees. Through simulation, they find that a heuristic taking both these factors into account can perform more efficiently than decentralized algorithms using only one of these sources of information. Finding the optimal way to combine location and degree information in decentralized search, and understanding the range of networks that are searchable under such optimal strategies, is an interesting direction for further research.

## 8. Conclusion

We have followed a particular strand of research running through the topic of complex networks, concerned with short paths and the ability of decentralized algorithms to find them. As suggested initially, the sequence of ideas here is characteristic of the flavor of research in this area: an experiment in the social sciences that highlights a fundamental and non-obvious property of networks (efficient searchability, in this case); a sequence of random graph models and accompanying analysis that seeks to capture this notion in a simple and stylized form; a set of measurements on large-scale network data that parallels the properties of the models, in some cases to a surprising extent; and a range of connections to further results and questions in algorithms, graph theory, and discrete probability.

To indicate some of the further directions in which research on this topic could proceed, we conclude with a list of open questions and issues related to small-world networks and decentralized search. Some of these questions have already come up implicitly in the discussion thus far, so one goal of this list is to collect a number of these questions in a single place. Other questions here, however, bring in issues that reach beyond the context of the earlier sections. And as with any list of open questions, we must mention a few caveats: the questions here take different forms, since some are concretely specified while other are more designed to suggest problems in need of a precise formulation; the questions are not independent, in that the answer to one might well suggest ways of approaching others; and several of the questions may well become more interesting if the underlying model or formulation is slightly varied or tweaked.

**1. Variation in node degrees.** As we discussed at the end of the previous section, decentralized search in models that combine wide variation in node degrees with some kind of spatial embedding is an interesting issue that is not well understood. Simsek and Jensen's study [58] of this issue left open the question of proving bounds on the efficiency of decentralized algorithms. For example, consider the $d$-dimensional grid-based model with exponent $\alpha$, and suppose that rather than constructing a fixed number of long-range contacts for each node, we draw the number of long-range contacts for each node $v$ independently from a given probability distribution. To be concrete, we could consider a distribution in which one selects $k$ long-range contacts with probability proportional to $k^{-\delta}$ for a constant $\delta$.

We now have a family of grid-based models parameterized by $\alpha$ and $\delta$, and we can study the performance of decentralized search algorithms that know not only the long-range contacts out of the current node, but also the degrees of the neighboring nodes. Decentralized selection of a neighbor for forwarding the message has a stochastic optimization aspect here, balancing the goal of forwarding to a node close to the target with the goal of forwarding to a high-degree node. We can now ask the general question of how the delivery time of decentralized algorithms varies in both $\alpha$ and $\delta$. Note that it is quite possible this question becomes more interesting if we vary the model so that long-range links are undirected; this way, a node with a large degree is both easy to find and also very useful once it is found. (In a directed version, a node with large out-degree may be relatively useless simply because it has low in-degree and so is unlikely to be found.)

**2. The case of $\alpha = 2d$.** In both the grid-based model and the related long-range percolation models, very little is known about the diameter of the graph when $\alpha$ is equal to twice the dimension. (It appears that a similar question arises in other versions of the group-based models from Section 5, when nodes form links with probability inversely proportional to the square of the size of the smallest group containing both of them.) Resolving the behavior of the diameter would shed light on this transitional point, which lies at the juncture between "small worlds" and "large worlds." This open

question also manifests itself in the gossip problem discussed in Section 7, where we noted that the transitional value $\alpha = 2d$ arises in distributed computing applications (see the discussion in [34], [54]).

**3. Paths of logarithmic length.** It would be interesting to know whether there is a decentralized algorithm in the $d$-dimensional grid-based model, at the "searchable exponent" $\alpha = d$, that could construct paths of length $O(\log n)$ while visiting only a polylogarithmic number of nodes. This would improve the result of Lebhar and Schabanel [42] to an asymptotically tight bound on path length.

**4. Small-world networks with an arbitrary base graph.** It would also be interesting to resolve the open problem of Fraigniaud [27] described in Section 7, formalizing the question of whether any graph can be turned into an efficiently searchable small world by appropriately adding long-range links[1].

**5. Extending the group-based model.** Theorem 5.2 on the group-based model contained a positive result generalizing the ones for grids and hierarchies, and it contained a general negative result for the case when long-range connection were "too long-range" (i.e. with exponent $\gamma < 1$). However, it does not fully generalize the results for grids and hierarchies, because there are set systems satisfying conditions (i), (ii), and (iii) of the theorem for which efficient decentralized search is possible even for exponents $\gamma > 1$. It would be interesting to find a variation on these three properties that still generalizes grids and hierarchies in a natural way, and for which $\gamma = 1$ is the unique exponent at which efficient decentralized search is possible.

**6. Multiple hierarchies.** Obtaining provable bounds for decentralized search in the "multiple hierarchies" model of Watts, Dodds, and Newman [63] is also an open question. Such results could form an interesting parallel with the findings they discovered through simulation. With some small modifications to the model of Watts et al., one can cast it in the group-based model of Section 5, and so it is entirely possible that progress on this question and the previous could be closely connected.

**7. The evolution of searchable networks.** The remaining questions have a more general flavor, where much of the challenge is the formalization of the underlying issue. To begin with, the current models supporting efficient decentralized search are essentially *static*, in that they describe how the underlying network is organized without suggesting how it might have evolved into this state. What kinds of growth processes or selective pressures might exist to cause networks to become more efficiently searchable? Interesting network evolution models addressing this question have been proposed by Clauset and Moore [19] and by Sandberg [56], both based on feedback mechanisms by which nodes repeatedly perform decentralized searches and

[1] Note added in proof: Fraigniaud, Lebhar, and Lotker have very recently announced a negative resolution of this question, constructing a family of graphs that cannot be turned into efficiently searchable small worlds by this process.

in the process partially "rewire" the network. Obtaining provable guarantees for these models, or variations on them, is an open question. A number of peer-to-peer file-sharing systems include similar feedback mechanisms, achieving good performance in practice. Freenet [18] is a good example of such a system, and the relationship of its feedback mechanism to the evolution of small-world networks is studied by Zhang et al. [67].

Game theory may provide another promising set of techniques for studying the evolution of small-world networks. A growing body of recent work has considered game-theoretic models of network formation, in which agents controlling nodes and edges interact strategically to construct a graph – the basic question is to understand what types of structures emerge when each agent is motivated by self-interest. For surveys of this area, see [5], [33], [65]. In the present case, it would be interesting to understand whether there are ways to define incentives such that the collective outcome of self-interested behavior would be a searchable small-world network.

**8. Decentralized search in the presence of incentives.** Game-theoretic notions can provide insight not just into the growth of a network, but also into the processes that operate on it. A topic of interest in peer-to-peer systems, as well as in the design of on-line communities, is the way in which the incentives offered to the members of the system influence the extent to which they are willing to forward queries and information. In the case of decentralized search, suppose that there is some utility associated with routing the message from the starting node to the target, and intermediate nodes behave strategically, demanding compensation for their participation in the construction of the path. How do results on decentralized path formation change when such behavior is incorporated into the model?

In [40], this question is made precise in a setting where the underlying network is a random tree, constructed via a branching process. It would be interesting to consider analogous issues in richer classes of networks.

**9. Reconstruction.** The networks we have considered here have all been embedded in some underlying "reference frame" – grids, hierarchies, or set systems – and most of our analysis has been predicated on a model in which the network is presented together with this embedding. This makes sense in many contexts; recall, for example, the discussion from Section 6 of network data explicitly embedded in Web topic directories [49], corporate hierarchies [1], or the geography of the U.S. [43]. In some cases, however, we may be presented with just the network itself, and the goal is to determine whether it has a natural embedding into a spatial or hierarchical structure, and to recover this embedding if it exists. For example, we may have data on communication within an organization, and the goal is to reconstruct the hierarchical structure under the assumption that the frequency of communication decreases according to a hierarchical model – or to reconstruct the positions of the nodes under the assumption that the frequency of communication decreases with distance according to a grid-based or rank-based model.

One can formulate many specific questions of this flavor. For example, given a network known to be generated by the grid-based model with a given exponent $\alpha$, can we approximately reconstruct the positions of the nodes on the grid? What if we are not told the exponent? Can we determine whether a given network was more likely to have been generated from a grid-based model with exponent $\alpha$ or $\alpha'$? Or what if there are multiple long-range contacts per node, and we are only shown the long-range edges, not the local edges? A parallel set of questions can be asked for the hierarchical model.

Questions of this type have been considered by Sandberg [55], who reports on the results of computational experiments but leaves open the problem of obtaining provable guarantees. Benjamini and Berger [11] pose related questions, including the problem of reconstructing the dimension $d$ of the underlying lattice when presented with a graph generated by long-range percolation on a finite piece of $\mathbb{Z}^d$.

**10. Comparing network datasets.** As we saw earlier, the models proposed in Sections 4 and 5 suggest a general perspective from which to analyze network datasets, by studying the way in which the density of links decays with increasing distance or increasing group size (e.g. [1], [43]). One could naturally use this style of analysis to compare related network datasets – for example taking the patterns of communication within $k$ different organizations (as Adamic and Adar did for the corporate lab they studied), and determining exponents $\gamma_1, \gamma_2, \ldots, \gamma_k$ for each such that the probability of a link between individuals $v$ and $w$ in a group of size $g$ scales approximately as $g^{-\gamma_i}$ in the $i^{\text{th}}$ organization. Differences among these exponents would suggest structural differences between the organizations at a global level – communication in some is more long-range, while in others it is more clustered at the low levels of the hierarchy. It would be interesting to understand whether these differences in turn were naturally reflected in other aspects of the organizations' behavior and performance.

More generally, large-scale social, technological, and information networks are sufficiently complex objects that the guiding principles provided by simple models seem crucial for our understanding of them. The perspective suggested here has offered one such collection of principles, highlighting in particular the ways in which these networks are intertwined with the spatial and organizational structures that they inhabit. One can hope that as we gather an increasing range of different perspectives, our understanding of complex networks will continue to deepen into a rich and informative theory.

## References

[1] Adamic, L., Adar, E., How to search a social network. *Social Networks* **27** (3) (2005), 187–203.

[2] Adamic, L. A., Lukose, R. M., Puniyani, A. R., Huberman, B. A., Search in Power-Law Networks. *Phys. Rev. E* **64** (2001), 46135–46143.

[3] Aizenman, M., Newman, C. M., Discontinuity of the Percolation Density in One-Dimensional $1/|x-y|^2$ Percolation Models. *Comm. Math. Phys.* **107** (1986), 611–647.

[4] Albert, R., Barabási, A.-L., Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** (2002), 47–97.

[5] Anshelevich, E., Network Design and Management with Strategic Agents. Ph.D. thesis, Cornell University, 2005.

[6] Aspnes, J., Shah, G., Distributed data structures for P2P systems. In *Theoretical and Algorithmic Aspects of Sensor, Ad Hoc Wireless and Peer-to-Peer Networks* (ed. by Jie Wu, ed.), CRC Press, Boca Raton, FL, 2005, 685–700.

[7] Assouad, P., Plongements lipschitziens dans $\mathbf{R}^n$. *Bull. Soc. Math. France* **111** (1983), 429–448.

[8] Barabási, A.-L., *Linked: the new science of networks.* Perseus, Cambridge, Mass., 2002.

[9] Barrière, L., Fraigniaud, P., Kranakis, E., Krizanc, D., Efficient Routing in Networks with Long Range Contacts. In *Distributed Computing*, Lecture Notes in Comput. Sci. 2180, Springer-Verlag, Berlin 2001, 270–284.

[10] Ben-Naim, Eli, Frauenfelder, Hans, Toroczkai, Zoltan (eds.), *Complex Networks*. Lecture Notes in Phys. 650, Springer-Verlag, Berlin 2004.

[11] Benjamini, I., Berger, N., The diameter of long-range percolation clusters on finite cycles. *Random Structures Algorithms* **19** (2001), 102–111.

[12] Benjamini, I., Kesten, H., Peres, Y., Schramm, O., Geometry of the uniform spanning forest: transitions in dimensions 4, 8, 12, . . . . *Ann. of Math.* (2) **160** (2004), 465–491.

[13] Biskup, M., On the scaling of the chemical distance in long range percolation models. *Ann. Probab.* **32** (2004), 2938–2977.

[14] Bollobás, B., *Random Graphs.* 2nd edition, Cambridge Stud. Adv. Math. 73, Cambridge University Press, Cambridge 2001.

[15] Bollobás, B., Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks* (ed. by S. Bornholdt, H. G. Schuster), Wiley/VCH, Weinheim 2003, 1–34.

[16] Bollobás, B., Chung, F. R. K., The diameter of a cycle plus a random matching. *SIAM J. Discrete Math.* **1** (1988), 328–333.

[17] Bollobás, B., de la Vega, W. F., The diameter of random regular graphs. *Combinatorica* **2** (1982), 125–134.

[18] Clarke, I., Sandberg, O., Wiley, B., Hong, T. W., Freenet: A Distributed Anonymous Information Storage and Retrieval System. In *Designing Privacy Enhancing Technologies*, Lecture Notes in Comput. Sci. 2009, Springer-Verlag, Berlin 2001, 46–66

[19] Clauset, A., Moore, C., How Do Networks Become Navigable? Preprint, 2003; arxiv.org, cond-mat/0309415.

[20] Coppersmith, D., Gamarnik, D., Sviridenko, M., The diameter of a long-range percolation graph. *Random Structures Algorithms* **21** (2002), 1–13.

[21] Demers, A., Greene, D., Hauser, C., Irish, W., Larson, J., Shenker, S., Sturgis, H., Swinehart, D., and Terry, D., Epidemic algorithms for replicated database maintenance. *ACM SIGOPS Operating Systems Review* **22** (1) (1988), 8–32.

[22] Dodds, P., Muhamad, R., Watts, D. J., An Experimental Study of Search in Global Social Networks. *Science* **301** (2003), 827.

[23] Dorogovtsev, S. N., Mendes, J. F. F., *Evolution of Networks: from biological networks to the Internet and WWW*. Oxford University Press, New York 2003.

[24] Duchon, P., Hanusse, N., Lebhar, E., Schabanel, N., Could any graph be turned into a small world? *Theoret. Comput. Sci.* **355** (1) (200), 96–103.

[25] Durrett, R., *Random Graph Dynamics.* Cambridge University Press, Cambridge 2006.

[26] Erdös, P., and Rényi, A., On the Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 17–61.

[27] Fraigniaud, P., A New Perspective on the Small-World Phenomenon: Greedy Routing in Tree-Decomposed Graphs. In *Proceedings of the 13th Annual European Symposium on Algorithms* (*ESA*), 2005.

[28] Fraigniaud, P., Gavoille, C., and Paul, C., Eclecticism shrinks even small worlds. In *Proceedings of the 23rd Annual Symposium on Principles of Distributed Computing*, ACM Press, New York 2004, 169–178.

[29] Guare, J., *Six Degrees of Separation: A Play.* Vintage Books, New York 1990.

[30] Gupta, A., Krauthgamer, R., and Lee, J. R., Bounded geometries, fractals, and low-distortion embeddings. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2003, 534–543.

[31] Higham, D., Greedy Pathlengths and Small World Graphs. University of Strathclyde Mathematics Research Report 08, 2002.

[32] Hong, T., Performance. In *Peer-to-Peer: Harnessing the Power of Disruptive Technologies* (ed. by A. Oram), O'Reilly and Associates, Sebastopol, CA, 2001, 203–241.

[33] Jackson, M., A Survey of Models of Network Formation: Stability and Efficiency. In *Group Formation in Economics; Networks, Clubs and Coalitions* (ed. by G. Demange and M. Wooders), Cambridge University Press, Cambridge 2004, 11–57.

[34] Kempe, D., Kleinberg, J., Demers, A., Spatial Gossip and Resource Location Protocols. In *Proceedings of the 33rd Annual Symposium on Theory of Computing*, ACM Press, New York 2001, 163–172.

[35] Killworth, P., and Bernard, H., Reverse small world experiment. *Social Networks* **1** (1978), 159–192.

[36] Kleinberg, J., Navigation in a Small World. *Nature* **406** (2000), 845.

[37] Kleinberg, J., The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd Annual Symposium on Theory of Computing*, ACM Press, New York 2000, 163–170.

[38] Kleinberg, J., Small-World Phenomena and the Dynamics of Information. In *Advances in Neural Information Processing Systems* (*NIPS*) **14** (2001), 431–438.

[39] Kleinberg, J., Lawrence, S., The Structure of the Web. *Science* **294** (2001), 1849–1850.

[40] Kleinberg, J., Raghavan, P., Query Incentive Networks. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2005.

[41] Kleinfeld, J., Could it be a Big World After All? The 'Six Degrees of Separation' Myth. *Society*, April 2002.

[42] Lebhar, E., Schabanel, N., Almost optimal decentralized routing in long-range contact networks. In *Proceedings of the International Colloquium on Automata, Languages, and Programming* (*ICALP*), 2004.

[43] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* **102** (2005), 11623–11628.

[44] Lua, E.-K., Crowcroft, J., Pias, M., Sharma, R., and Lim, S., A Survey and Comparison of Peer-to-Peer Overlay Network Schemes. *IEEE Communications Surveys and Tutorials* **7** (2005), 72–93.

[45] Malkhi, D., Naor, M., Ratajczak, D., Viceroy: a scalable and dynamic emulation of the butterfly. In *Proceedings of the 21st Annual Symposium on Principles of Distributed Computing*, ACM Press, New York 2002, 183–192.

[46] Manku, G. S., Bawa, M., Raghavan, P., Symphony: Distributed hashing in a small world. In *Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems*, USENIX, 2003, 127–140.

[47] Manku, G., Naor, M., and Wieder, U., Know Thy Neighbor's Neighbor: The Power of Lookahead in Randomized P2P Networks. In *Proceedings of the 36th Annual Symposium on Theory of Computing*, ACM Press, New York 2004, 54–63.

[48] Martel, C., Nguyen, V., Analyzing Kleinberg's (and other) small-world models. In *Proceedings of the 23rd Annual Symposium on Principles of Distributed Computing*, ACM Press, New York 2004, 179–188.

[49] Menczer, F., Growing and Navigating the Small World Web by Local Content. *Proc. Natl. Acad. Sci. USA* **99** (22) (2002), 14014–14019.

[50] Milgram, S., The small world problem. *Psychology Today* **1** (1967), 60–67.

[51] Newman, C. M., Schulman, L. S., One Dimensional $1/|j-i|^s$ Percolation Models: The Existence of a Transition for $s \leq 2$. *Comm. Math. Phys.* **104** (1986), 547–571.

[52] Newman, M. E. J., The structure and function of complex networks. *SIAM Review* **45** (2003), 167–256.

[53] Nguyen, V., and Martel, C., Analyzing and characterizing small-world graphs. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, Philadelphia, PA, 2005, 311–320.

[54] van Renesse, R., Birman, K. P., Vogels, W., Astrolabe: A robust and scalable technology for distributed system monitoring, management, and data mining. *ACM Trans. Computer Sys.* **21** (2003), 164–206.

[55] Sandberg, O., Distributed Routing in Small-World Networks. In *Proceedings of the 8th Workshop on Algorithm Engineering and Experiments*, 2006.

[56] Sandberg, O., Searching a Small World. Licentiate thesis, Chalmers University, 2005.

[57] Schulman, L. S., Long-range percolation in one dimension. *J. Phys. A* **16** (17) (1986), L639–L641.

[58] Simsek, O., and Jensen, D., Decentralized search in networks using homophily and degree disparity. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 2005.

[59] Slivkins, A., Distance Estimation and Object Location via Rings of Neighbors. In *Proceedings of the 24rd Annual Symposium on Principles of Distributed Computing*, ACM Press, New York 2005, 41–50.

[60] Strogatz, S., Exploring complex networks. *Nature* **410** (2001), 268.

[61] Travers, J., and Milgram, S., An experimental study of the small world problem. *Sociometry* **32** (1969), 425–443.

[62] Watts, Duncan J., *Six Degrees: The Science of a Connected Age*. W. W. Norton, Scranton, PA, 2003.

[63] Watts, D. J., Dodds, P. S., Newman, M. E. J., Identity and Search in Social Networks. *Science* **296** (2002), 1302–1305.

[64] Watts, D. J. and Strogatz, S. H., Collective dynamics of 'small-world' networks. *Nature* **393** (1998), 440–442.

[65] Wexler, T., Pricing Games with Selfish Users. Ph.D. thesis, Cornell University, 2005.

[66] Zeng, J., Hsu, W.-J., Wang, J., Near Optimal Routing in a Small-World Network with Augmented Local Awareness. In *Proceedings of the 3rd International Symposium on Parallel and Distributed Processing and Applications*, 2005.

[67] Zhang, H., Goel, A., Govindan, R., Using the Small-World Model to Improve Freenet Performance. In *Proc. IEEE Infocom*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2002.

Department of Computer Science, Cornell University, Ithaca, NY 14853, U.S.A.

# On expander graphs and connectivity in small space

Omer Reingold*

**Abstract.** This presentation is aimed to communicate a recently found deterministic algorithm for determining connectivity in undirected graphs [40]. This algorithm uses the minimal amount of memory possible, up to a constant factor. Specifically, the algorithm's memory is comparable to that needed to store the name of a *single* vertex of the graph (i.e., it is logarithmic in the size of the graph).

Our algorithm also implies a deterministic, short (i.e. of polynomial length), universal sequence of steps which explores all the edges of every regular undirected graph. Such a sequence will get one out of every maze, and through the streets of every city. More formally we give universal exploration sequences for arbitrary graphs and universal traversal sequences for graphs with some natural restriction on their labelling. Both sequences are constructible with logarithmic memory and are thus only polynomially long.

To obtain this algorithm, we give a method to transform (using small memory), an arbitrary connected undirected graph into an expander graph (which is a sparse but highly connected graph).

**Mathematics Subject Classification (2000).** Primary 05C40; Secondary 68Q15.

**Keywords.** Graph connectivity, expander graphs, randomness in computation, symmetric logspace (SL), pseudorandomness.

## 1. Introduction

An undirected graph is a pair of finite sets $G = (V, E)$ where $V$ is the set of *vertices* and $E$ is the set of *edges*. An edge is simply a pair of vertices (we say that the edge is adjacent to these two vertices and the two vertices are adjacent to it). Special cases of graphs that may be useful to visualize are mazes and road maps where a vertex corresponds to an intersection and an edge directly connects the two vertices adjacent to it.[1] Given as input an undirected graph $G$ and two vertices $s$ and $t$, the undirected st-connectivity (denoted USTCON) problem is to decide whether or not the two vertices are connected by a path (i.e. a sequence of edges) in $G$ . This fundamental combinatorial problem has received a lot of attention in the last few decades and was studied in a large variety of computational models. It is a basic building block for more complex graph algorithms and is complete[2] for an interesting class

*Incumbent of the Walter and Elise Haas Career Development Chair, Research supported by US-Israel Binational Science Foundation Grant 2002246.

[1] In a directed graph an edge is an ordered pair of vertices and we can therefore think of an edge as a unidirectional road going from the first vertex to the second vertex.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of computational problems known as SL (these are problems solvable by symmetric, non-deterministic, log-space computations [25]). A few interesting examples of problems in SL are deciding if a graph is bipartite (equivalently if it is 2-colorable), if a bounded degree graph is planar and identifying chordal graphs, interval graphs, split graphs and permutation graphs (see [8] for a recent study of SL and quite a few of its complete problems).

The time complexity of USTCON is well understood as basic search algorithms, particularly breadth-first search (BFS) and depth-first search (DFS), are capable of solving USTCON in linear time. In fact, these algorithms apply to the more complex problem of st-connectivity in directed graphs, denoted STCON (which is complete for the class NL of non-deterministic log-space computations). Unfortunately, the space required to run these algorithms is linear as well. In a recent paper [40] we resolve the space (memory) complexity of USTCON, up to a constant factor, by presenting a log-space (polynomial-time) algorithm for solving it (our algorithm also solves the corresponding search problem, of finding a path from $s$ to $t$ if such a path exists). In this presentation (which is in large part borrowed from [40]) we will discuss this algorithm. We will also discuss the connection of this question to the possible tradeoff between two resources of computation: randomness and memory. Finally, we will discuss explicit (and space efficient) constructions of fascinating combinatorial objects known as universal traversal and universal exploration sequences. Loosely, these are short sequences of simple directions that lead a walk through all of the edges of any graph of an appropriate size. Our main technical tool is borrowed from a combinatorial construction of expander graphs due to Reingold, Vadhan and Wigderson [43]. We will further elaborate on the connection to expander graphs.

**Background.** Given the inefficiency in terms of memory of BFS and DFS, Savitch's [47] introduced an algorithm which solves STCON in space $\log^2(\cdot)$ (and super-polynomial time). Major progress in understanding the space complexity of USTCON was made by Aleliunas, Karp, Lipton, Lovász, and Rackoff [2], who gave a *randomized* log-space algorithm for the problem. Specifically, they showed that a random walk (a path that selects a uniform edge at each step) starting from an arbitrary vertex of any connected undirected graph will visit all the vertices of the graph in polynomial number of steps. Therefore, the algorithm can perform a random walk starting from $s$ and verify that it reaches $t$ within the specified polynomial number of steps. Essentially all that the algorithm needs to remember is the name of the current vertex and a counter for the number of steps already taken. With this result we get the following view of space complexity classes: $\mathrm{L} \subseteq \mathrm{SL} \subseteq \mathrm{RL} \subseteq \mathrm{NL} \subseteq \mathrm{L}^2$ (where RL is the class of problems that can be decided by randomized log-space algorithms with one-sided error and $\mathrm{L}^c$ is the class of problems that can be decided deterministically in space $\log^c(\cdot)$).

[2] A complete problem for a class is such that any other problem "efficiently reduces" to it. Therefore, an algorithm for this problem implies an algorithm (which is almost as efficient) for any other problem in the class.

The existence of a randomized log-space algorithm for USTCON puts this problem in the context of derandomization. Can this randomized algorithm be derandomized without substantial increase in space? Furthermore, the study of the space complexity of USTCON has gained additional motivation as an important test case for understanding the tradeoff between two central resources of computations, namely between memory space and randomness. Particularly, a natural goal on the way to proving RL = L is to prove that USTCON $\in$ L, as USTCON is undoubtedly one of the most interesting problems in RL.

Following [2], most of the progress on the space complexity of USTCON indeed relied on the tools of derandomization. In particular, this line of work greatly benefited from the development of pseudorandom generators that fool space-bounded algorithms [1], [10], [33], [19] and it progressed concurrently with the study of the L vs. RL problem. Another very influential notion, introduced by Stephen Cook in the late 70s, is that of a universal-traversal sequence. Loosely, this is a fixed sequence of directions that guides a *deterministic* walk through all of the vertices of any connected graph of the appropriate size (see further discussion below).

While Nisan's space-bounded generator [33], did not directly imply a more space efficient USTCON algorithm it did imply quasi-polynomially-long, universal-traversal sequences, constructible in space $\log^2(\cdot)$. These were extremely instrumental in the work of Nisan, Szemeredi and Wigderson [34] who showed that USTCON $\in \mathrm{L}^{3/2}$ – The first improvement over Savitch's algorithm in terms of space (limited of course to the case of undirected graphs). Using different methods, but still heavily relying on [33], Saks and Zhou [46] showed that *every* RL *problem* is also in $\mathrm{L}^{3/2}$ (their result in fact generalizes to randomized algorithms with two-sided error). Relying on the techniques of both [34] and [46], Armoni, et. al. [9] showed that USTCON $\in \mathrm{L}^{4/3}$. Their USTCON algorithm was the most space-efficient one previous to this work. We note that the most space-efficient *polynomial-time* algorithm for USTCON previously known was Nisan's [32], which still required space $\log^2(\cdot)$. Independent of our work (and using different techniques), Trifonov [49] has presented an $O(\log n \log\log n)$-space, deterministic algorithm for USTCON.

**Our approach.** The essence of our algorithm is in the following very natural approach: If you want to solve a connectivity problem on your input graph, first *improve its connectivity*. In other words, transform your input graph (or rather, each one of its connected components), into an expander.[3] We will also insist on the final graph being constant degree (i.e., every vertex is adjacent to a constant number of edges). Once the connected component of $s$ is a constant-degree expander, then it is trivial to decide if $s$ and $t$ are connected: Since expander graphs have logarithmic diameter,

[3]The exact definition of expander graphs is less important for now, and the following description could be understood by viewing expanders as graphs with very strong connectivity properties. Still, for the knowledgable reader, the particular measure that seems the most convenient to work with is the second eigenvalue (in absolute value) of the adjacency matrix of the graph (we will only need to work with regular graphs). It may however be that other, more combinatorial, measures will also do (see [41] for a more detailed discussion).

it is enough to enumerate all logarithmically long paths starting with $s$ and to see if one of these paths visits $t$. Since the degree is constant, the number of such paths is polynomial and they can easily be enumerated in log space.

How can we turn an arbitrary graph into an expander? First, we note that every connected, non-bipartite, graph can be thought of as an expander with very small (but non-negligible) expansion. Consider for example an arbitrary connected graph with self-loops added to each one of its vertices. The number of neighbors of every strict subset of the vertices is larger than its size by at least one. In this respect, the graph can be thought of as expanding by a factor $1 + 1/N$ (where $N$ is the total number of vertices in the graph). Now, a very natural operation that improves the expansion of the graph is powering. The $k^{\text{th}}$ power of $G$ contains an edge between two vertices $v$ and $w$ for every path of length $k$ in $G$. Formally, it can be shown that by taking some polynomial power of any connected non-bipartite graph (equivalently, by repeatedly squaring the graph logarithmic number of times), it will indeed turn into an expander.

The down side of powering is of course that it increases the degree of the graph. Taking a polynomial or any non-constant power is prohibited if we want to maintain constant degree. Fortunately, there exist operations that can counter this problem. Consider for example the replacement product of a $D$-regular graph $G$ with a $d$-regular graph $H$ on $D$ vertices (with $d \ll D$). This can be loosely defined as follows: Each vertex $v$ of $G$ is replaced with a "copy" $H_v$ of $H$. Each of the $D$ vertices of $H_v$ is connected to its neighbors in $H_v$ but also to one vertex in $H_w$, where $(v, w)$ is one of the $D$ edges going out of $v$ in $G$. The degree in the product graph is $d + 1$ (which is smaller than $D$). Therefore, this operation can transform a graph $G$ into a new graph (the product of $G$ and $H$) of smaller degree. It turns out that if $H$ is a "good enough" expander, the expansion of the resulting graph is "not worse by much" than the expansion of $G$. Formal statements to this affect were proven by Reingold, Vadhan and Wigderson [43] for both the replacement product and the zig-zag product, introduced there. Independently, Martin and Randall [30], building on previous work of Madras and Randall [27], proved a decomposition theorem for Markov chains that also implies that the replacement product preserves expansion.

Given the discussion above, we are ready to informally describe our USTCON algorithm. First, turn the input graph into a constant-degree, regular graph with each connected component being non-bipartite (this step is very easy). Then, the main transformation turns each connected component of the graph, in logarithmic number of phases, into an expander. Each phase starts by raising the current graph to some constant power and then reducing the degree back via a replacement or a zig-zag product with a constant-size expander. We argue that each phase enhances the expansion at least as well as squaring the graph would, and *without the disadvantage of increasing the degree*. Finally, all that is left is to solve USTCON on the resulting graph (which is easy as the diameter of each connected component is only logarithmic).

To conclude that USTCON $\in$ L, we need to argue that all of the above can be done in logarithmic space, which easily reduces to showing that the main transformation can be carried out in logarithmic space. For that, consider the graph $G_i$ obtained after $i$

phases of the transformation. We note that a step on $G_i$ (i.e., evaluating the $j$'th neighbor of some vertex $v$ in $G_i$) is composed of a constant number of operations that are either a step on the graph $G_{i-1}$ from the previous phase or an operation that only requires a constant amount of memory. As the memory for each of these operations can be freed after it is performed, the memory for carrying out a step on $G_i$ is only larger by an additive constant than the memory for carrying out a step on $G_{i-1}$. This implies that the entire transformation is indeed log space.

**Universal traversal sequences** While universal-traversal sequences were introduced as a way for proving USTCON $\in$ L, these are interesting combinatorial objects in their own right. A universal-traversal sequence for $D$-regular graphs on $N$-vertices, is a sequence of edge labels in $\{1, \ldots, D\}$ such that for every such graph, for every labelling of its edges, and for every start vertex, the *deterministic* walk defined by these labels (where in the $i$'th step we take the edge labeled by the $i$'th element of the sequence), visits all of the vertices of the graph. Aleliunas et. al. [2] showed that polynomial-length universal-traversal sequence exists, and in fact almost every sequence of the appropriate length will do. We are interested in obtaining a polynomially-long, universal-traversal sequence that is *constructible in logarithmic space* (even less explicit sequences may still be very interesting). This is again a derandomization problem. Namely, can we derandomize the probabilistic construction of universal-traversal sequences?

Explicit constructions of polynomially-long universal-traversal sequences are only known for extremely limited classes of graphs. Even for expander graphs, such sequences are only known when the edges are "consistently labelled" [18] (this means that the labels of all edges that lead to any particular vertex are distinct). It is therefore not very surprising that our algorithm on its own does not imply full fledged universal-traversal sequences. Still, our algorithm can be shown to imply a very local, and quite oblivious, deterministic procedure for exploring a graph. We can think of our algorithm as maintaining a single pebble, that is placed on the *edges* of the graph. The pebble is moved either from one side of the edge to another, or between different edges that are adjacent to the same vertex (say to the next or to the previous edge). As with universal-traversal sequences, the fixed sequence of instructions is good for every graph, for every labelling of its edges, and for any starting point on the graph. The only difference from universal-traversal sequences is that the pebble here is placed on the edges rather than on the vertices of the graph. In particular, we get polynomially-long, universal-*exploration* sequences for all undirected graphs. In universal-exploration sequences, introduced by Koucky [23], the elements of the sequence are not interpreted as absolute edge-labels but rather as offsets from the previous edge that was traversed. In terms of traversal sequences, our algorithm implies a polynomially-long, universal-traversal sequence that is *constructible in logarithmic space* under some restrictions on the labelling. These restrictions were relaxed in a subsequent work [41] to be identical to those of [18]. For more details see Section 5.

**More on previous work** Graph connectivity problems and space-bounded derandomization are the focus of a vast and diverse body of research. The scope of this paper only allows for an extremely partial discussion of this area. Some very beautiful and influential research (as many of the papers mentioned above) is only briefly touched upon, other areas will not be discussed at all (examples include, time-space tradeoffs for deterministic and randomized connectivity algorithms, restricted constructions of universal traversal sequences, and analysis of connectivity in many other computational models). Insightful, though somewhat outdated, surveys on these topics were given by Wigderson [50] and by Saks [45]. Useful discussion and pointers were also given by Koucky [24]. We continue here by mentioning a few of the most related previous results (most of which are subsumed by the results of this paper). A more technical comparison with some previous work appears in Section 6.

Following Aleliunas et. al. [2], Borodin et. al. [12] gave a *zero-error*, randomized, log-space algorithm for USTCON. An upper bound of different nature on SL was given by Karchmer and Wigderson [21], who showed $\mathrm{SL} \subseteq \oplus\mathrm{L}$.

Nisan and Ta-Shma [35] showed that SL is closed under complement, thus collapsing the "symmetric log-space hierarchies" of both Reif [39] and Ben Asher et. al. [11], and putting some very interesting problems into SL. To give just one example, the planarity of bounded-degree undirected graphs was placed in SL as a corollary (we refer again to [8] for a list of SL-complete problems).

A research direction initiated by Ajtai et. al. [1], and continued with Nisan and Zuckerman [36] is to fully derandomize (i.e., to put in L) $\log n$-space computations that use fewer than $n$ random bits (poly $\log n$ bits in the case of [36]). Raz and Reingold [38] showed how to derandomize $2^{\sqrt{\log n}}$ bits for subclasses of RL. One of their main applications can be viewed as derandomizing $2^{\sqrt{\log n}}$ bits for SL. It is interesting to note (and personally gratifying to the author) that the techniques of [38] played a major roll in the definition of the zig-zag product and with this work found their way back to the study of space-bounded derandomization.

Goldreich and Wigderson [17] gave an algorithm that on all but a tiny fraction of the graphs, evaluates USTCON correctly (and on the rest of the graphs outputs an error message).

Based on rather relaxed *computational hardness assumptions*, Klivans and van Melkebeek [22] proved both that $\mathrm{RL} = \mathrm{L}$ and that efficiently constructible, polynomial length, universal traversal sequences exist.

## 2. Preliminaries

This section discusses various aspects of graphs: their representation, eigenvalue expansion, graph powering, and two graph products (the replacement product and the zig-zag product). The definitions and notation used here are borrowed directly from [43].

**2.1. Graphs representations.** There are several standard representations of graphs. Fortunately, there exist log-space transformations between natural representations. Thus, the space complexity of USTCON is to a large extent independent of the representation of the input graph.

When discussing the eigenvalue expansion of a graph, we will consider its adjacency matrix. That is, the matrix whose (nonnegative, integral) entry $(u, v)$ equals to the number of edges that go from vertex $u$ to vertex $v$. Note that this representation allows graphs with self loops and parallel edges (and indeed such graphs may be generated by our algorithm). A graph is *undirected* iff its adjacency matrix is symmetric (implying that for every edge from $u$ to $v$ there is an edge from $v$ to $u$). It is *$D$-regular* if the sum of entries in each row (and column) is $D$ (so exactly $D$ edges are incident to every vertex).

Let $G$ be a $D$-regular undirected graph on $N$ vertices. When considering a walk on $G$, we would like to assume that the edges leaving each vertex of $G$ are labeled from 1 to $D$ in some arbitrary, but fixed, way. We can then talk about the $i$'th edge incident to a vertex $v$, and similarly about the $i$'th neighbor of $v$. A central insight of [43] is that when taking a step on a graph from vertex $v$ to vertex $w$, it may be useful to keep track of the edge traversed to get to $w$ (rather than just remembering that we are now at $w$). This gave rise to a new representation of graphs through the following *permutation* on pairs of vertex name and edge label:

**Definition 2.1.** For a $D$-regular undirected graph $G$, the *rotation map* $\mathrm{Rot}_G : [N] \times [D] \to [N] \times [D]$ is defined as follows: $\mathrm{Rot}_G(v, i) = (w, j)$ if the $i$'th edge incident to $v$ leads to $w$, and this edge is the $j$'th edge incident to $w$.

Rotation maps will indeed be the representation of choice for this work. Specifically, the first step of our algorithm will be to transform the input graph into a regular one specified by its rotation map (in particular, this step will give labels to the edges of the graph).

**2.2. Eigenvalue expansion and st-connectivity for expanders.** Expanders are sparse graphs which are nevertheless highly connected. The strong connectivity properties of expanders make them very desirable in our context. Specifically, since the diameter of expander graphs is only logarithmically long, there is a trivial log-space algorithm for finding paths between vertices in constant-degree expanders. The particular formalization of expanders used in this paper is the (algebraic) characterization based on the spectral gap of their adjacency matrix. Namely, the gap between the first and second eigenvalues of the (normalized) adjacency matrix.

The *normalized adjacency matrix* $M$ of a $D$-regular undirected graph $G$, is the adjacency matrix of $G$ divided by $D$. In terms of the rotation map, we have:

$$M_{u,v} = \frac{1}{D} \cdot \left| \{(i, j) \in [D]^2 : \mathrm{Rot}_G(u, i) = (v, j)\} \right|.$$

$M$ is simply the transition probability matrix of a random walk on $G$. By the $D$-regularity of $G$, the all-1's vector $1_N = (1, 1, \ldots, 1) \in \mathbb{R}^N$ is an eigenvector of $M$ of

eigenvalue 1. It turns out that all the other eigenvalues of $M$ have absolute value at most 1. We denote by $\lambda(G)$, the *second largest eigenvalue* (in absolute value) of $G$'s normalized adjacency matrix. We refer to a $D$-regular undirected graph $G$ on $N$ vertices such that $\lambda(G) \leq \lambda$ as an $(N, D, \lambda)$-*graph*. It is well-known that the second largest eigenvalue of $G$ is a good measure of $G$'s expansion properties. In particular, it was shown by Tanner [48] and Alon and Milman [5] that second-eigenvalue expansion implies (and is in fact equivalent [3]) to the standard notion of *vertex expansion*. In particular, for every $\lambda < 1$ there exists $\varepsilon > 0$ such that for every $(N, D, \lambda)$-graph $G$ and for any set $S$ of at most half the vertices in $G$, at least $(1 + \varepsilon) \cdot |S|$ vertices of $G$ are connected by an edge to some vertex in $S$. This immediately implies that $G$ has a logarithmic diameter:

**Proposition 2.2.** *Let $\lambda < 1$ be some constant. Then for every $(N, D, \lambda)$-graph $G$ and any two vertices $s$ and $t$ in $G$, there exists a path of length $O(\log N)$ that connects $s$ to $t$.*

*Proof.* By the vertex expansion of $G$, for some $\ell = O(\log N)$ both s and $t$ have more than $N/2$ vertices of distance at most $\ell$ from them in $G$. Therefore, there exists a vertex $v$ that is of distance at most $\ell$ from both $s$ and $t$. □

We can therefore conclude that st-connectivity in constant-degree expanders can be solved in log-space:

**Proposition 2.3.** *Let $\lambda < 1$ be some constant. Then there exists a space $O(\log D \cdot \log N)$ algorithm $\mathcal{A}$ such that when a $D$-regular undirected graph $G$ on $N$ vertices is given to $\mathcal{A}$ as input, the following hold:*

1. *If $s$ and $t$ are in the same connected component and this component is an $(N', D, \lambda)$-graph then $\mathcal{A}$ outputs 'connected'.*
2. *If $\mathcal{A}$ outputs 'connected' then $s$ and $t$ are indeed in the same connected component.*

*Proof.* The algorithm $\mathcal{A}$ simply enumerates all $D^\ell$ paths of length $\ell = O(\log N)$ from $s$. (Where the leading constant in the big-$O$ notation depends on $\lambda$ as in Proposition 2.2.) The algorithm $\mathcal{A}$ outputs 'connected' if and only if at least one of these paths encounters $t$.

Following any particular path from $s$ of length $\ell$ requires space $O(\log N)$, (when given as input the sequence of $\ell$ edge labels in $[D] = \{1, 2, \ldots D\}$ traversed by this path). Enumerating all these $D^\ell$ paths requires space $O(\log D \cdot \log N)$. By Proposition 2.2, in case (1), $s$ and $t$ are of distance at most $\ell$ of each other and $\mathcal{A}$ will indeed find a path from $s$ to $t$ and will output 'connected'. On the other hand, $\mathcal{A}$ never outputs 'connected' unless it finds a path from $s$ to $t$, implying (2). □

Using the probabilistic method, Pinsker [37] showed that most 3-regular graphs are expanders (in the sense of vertex expansion), and this result was extended to eigenvalue

bounds in [3], [13], [15], [14]. Various *explicit* families of constant-degree expanders, some with optimal tradeoff between degree and expansion, were given in literature (cf. [28], [16], [20], [5], [4], [26], [29], [31], [43]). Our algorithm will employ a single constant size expander with rather weak parameters. This expander can be obtained by exhaustive search or by any of the explicit constructions mentioned above. In fact, one can use simpler explicit constructions than the ones given above, as we can afford a rather large degree (with respect to the number of vertices), rather than a constant degree. An example of a simpler construction that would suffice is the one given by Alon and Roichman [6], (see also related discussions in [43] regarding their "base graph").

**Proposition 2.4.** *There exists some constant $D_e$ and a $((D_e)^{16}, D_e, 1/2)$-graph.*

Finally, a key fact for our algorithm is that every connected, non-bipartite graph has a spectral gap which is at least inverse polynomial in the size of the graph (recall that a graph is non-bipartite if there is no partition of the vertices such that all the edges go between the two sides of the partition).

**Lemma 2.5** ([7]). *For every $D$-regular, connected, non-bipartite graph $G$ on $[N]$ it holds that $\lambda(G) \leq 1 - 1/DN^2$.*

**2.3. Powering.** Our main transformation will take a graph and transform each one of its connected components (that in itself will be a connected, non-bipartite graph), into a constant degree expander. If we ignore the requirement that the graph remains constant degree, a simple way of amplifying the (inverse polynomial) spectral gap of a graph is by powering.

**Definition 2.6.** Let $G$ be a $D$-regular multigraph on $[N]$ given by rotation map $\mathrm{Rot}_G$. The *$t$'th power* of $G$ is the $D^t$-regular graph $G^t$ whose rotation map is given by $\mathrm{Rot}_{G^t}(v_0, (a_1, a_2, \ldots, a_t)) = (v_t, (b_t, b_{t-1}, \ldots, b_1))$, where these values are computed via the rule $(v_i, b_i) = \mathrm{Rot}_G(v_{i-1}, a_i)$.

**Proposition 2.7.** *If $G$ is an $(N, D, \lambda)$-graph, then $G^t$ is an $(N, D^t, \lambda^t)$-graph.*

*Proof.* The normalized adjacency matrix of $G^t$ is the $t$'th power of the normalized adjacency matrix of $G$, so all the eigenvalues also get raised to the $t$'th power. □

**2.4. Two graph products.** While taking a power of a graph reduces its second eigenvalue, it also increases its degree. As we are interested in producing constant-degree graphs, we need a complementing operation that reduces the degree of a graph without harming its expansion by too much. We now discuss two graph products that are capable of doing exactly that.

The first is the very natural product, known as the *replacement product*. Assume that $G$ is a $D$-regular graph on $[N]$ and $H$ is a $d$-regular graph on $[D]$ (where $d$ is significantly smaller than $D$). Very intuitively, the replacement product of the two

graphs is defined as follows: Each vertex $v$ of $G$ is replaced with a "copy" $H_v$ of $H$. Each of the $D$ vertices of $H_v$ is connected to its neighbors in $H_v$ but also to one vertex in $H_w$, where $(v, w)$ is one of the $D$ edges going out of $v$ in $G$. The degree in the product graph is $d + 1$ (which is smaller than $D$).[4] A second, slightly more evolved, product introduced by Reingold, Vadhan and Wigderson [43], is the *zig-zag graph product*. Here too we replace each vertex $v$ of $G$ with a "copy" $H_v$ of $H$. However, the edges of the zig-zag product of $G$ and $H$ correspond to a subset of the paths of length three in the replacement product of these graphs[5] (see formal definition below). The degree of the product graph here is $d^2$ (which should still be thought of as significantly smaller that $D$).

It is immediate from their definition, that both products can transform a graph $G$ to a new graph (the product of $G$ and $H$) of smaller degree. As discussed in the introduction, it was previously shown [43], [30] that if $H$ is a "good enough" expander, then the expansion of the resulting graph is "not worse by much" than the expansion of $G$ (see formal statement below for the zig-zag product). Either one of these products can be used in our USTCON algorithm (with some variation in the parameters). We find it more convenient to work with the zig-zag product (even though it is a bit more involved), hence we proceed by formally defining it.

**Definition 2.8** ([43]). If $G$ is a $D$-regular graph on $[N]$ with rotation map $\mathrm{Rot}_G$ and $H$ is a $d$-regular graph on $[D]$ with rotation map $\mathrm{Rot}_H$, then their *zig-zag product* $G ⓩ H$ is defined to be the $d^2$-regular graph on $[N] \times [D]$ whose rotation map $\mathrm{Rot}_{G ⓩ H}$ is as follows (see Figure 1 for an illustration):

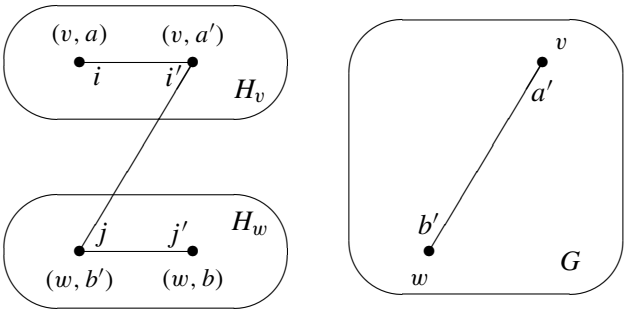



Figure 1. On the left – an edge of the zig-zag product is composed of three steps: a "short step" (in $H_v$), a "big step" (between $H_v$ and $H_w$ which corresponds to an edge of $G$ between $v$ and $w$), and a final "small step" (in $H_w$). The values $i$ ,$i'$, $j$ and $j'$ are labels of edges of $H$ (going out of the $H$ vertices $a$, $a'$, $b'$ and $b$ respectively). On the right – the projection of these steps on the graph $G$ (which corresponds to the middle step specified by $(w, b') = \mathrm{Rot}_G(v, a')$).

[4] Sometimes it is better to consider the *balanced* replacement product, where every edge in $G$ is taken $d$ times in parallel. The degree of the product graph in this case is $2d$ instead of $d + 1$.

[5] Those length three paths that are composed of a "short edge" (an edge inside one of the copies $H_v$), a "long edge" (one that corresponds to an edge of $G$), and finally one additional "short edge".

**$\mathbf{Rot}_{G ⓩ H}((v, a), (i, j))$:**

1. Let $(a', i') = \mathrm{Rot}_H(a, i)$.
2. Let $(w, b') = \mathrm{Rot}_G(v, a')$.
3. Let $(b, j') = \mathrm{Rot}_H(b', j)$.
4. Output $((w, b), (j', i'))$.

In [43], $\lambda(G ⓩ H)$ was bounded as a function of $\lambda(G)$ and $\lambda(H)$. The interesting case there was when both $\lambda(G)$ and $\lambda(H)$ were small constants (and in fact, $\lambda(G)$ is significantly smaller than $\lambda(H)$). In our context, $\lambda(H)$ will indeed be a small constant but $G$ may have an extremely small spectral gap (recall that the spectral gap of $G$ is $1 - \lambda(G)$). In this case, we want the spectral gap of $G ⓩ H$ to be roughly the same as that of $G$ (i.e., smaller by at most a constant factor). It turns out that the stronger bound on $\lambda(G ⓩ H)$, given in [43] implies a useful bound also in this case. We note that a simpler proof for the sort of bound on the zig-zag product we need is given in [41] (in a more general setting than the one considered in [43]).

**Theorem 2.9** ([43]). *If $G$ is an $(N, D, \lambda)$-graph and $H$ is a $(D, d, \alpha)$-graph, then $G ⓩ H$ is a $(N \cdot D, d^2, f(\lambda, \alpha))$-graph, where*

$$f(\lambda, \alpha) = \frac{1}{2}(1 - \alpha^2)\lambda + \frac{1}{2}\sqrt{(1 - \alpha^2)^2\lambda^2 + 4\alpha^2}.$$

As a simple corollary, we have that the spectral gap of $G ⓩ H$ is smaller than that of $G$ by a factor that only depends on $\lambda(H)$.

**Corollary 2.10.** *If $G$ is an $(N, D, \lambda)$-graph and $H$ is a $(D, d, \alpha)$-graph, then*

$$1 - \lambda(G ⓩ H) \geq \frac{1}{2}(1 - \alpha^2) \cdot (1 - \lambda).$$

*Proof.* Since $\lambda \leq 1$ we have that

$$\frac{1}{2}\sqrt{(1 - \alpha^2)^2\lambda^2 + 4\alpha^2} \leq \frac{1}{2}\sqrt{(1 - \alpha^2)^2 + 4\alpha^2} = \frac{1}{2}(1 + \alpha^2) = 1 - \frac{1}{2}(1 - \alpha^2).$$

Therefore, $f(\lambda, \alpha)$ from Theorem 2.9 satisfies $f(\lambda, \alpha) \leq 1 - \frac{1}{2}(1 - \alpha^2)(1 - \lambda)$. □

## 3. Transforming graphs into expanders

This section gives a log-space transformation that essentially turns each one of the connected components of a graph into an expander. This is the main part of our USTCON algorithm.

**Definition 3.1** (Main transformation). On input $G$ and $H$, where $G$ is a $D^{16}$-regular graph on $[N]$ and $H$ is a $D$-regular graph on $[D^{16}]$, both given by their rotation maps, the transformation $\mathcal{T}$ outputs the rotation map of a graph $G_\ell$ defined as follows:

- Set $\ell$ to be the smallest integer such that $(1 - 1/DN^2)^{2^\ell} < 1/2$.
- Set $G_0$ to equal $G$, and for $i > 0$ define $G_i$ recursively by the rule:

$$G_i = (G_{i-1} \textcircled{z} H)^8.$$

Denote by $\mathcal{T}_i(G, H)$ the graph $G_i$, and $\mathcal{T}(G, H) = G_\ell$

Note that by the basic properties of powering and the zig-zag product, it follows inductively that each $G_i$ is a $D^{16}$-regular graph over $[N] \times ([D^{16}])^i$. In particular, the zig-zag product of $G_i$ and $H$ is well defined. In addition, if $D$ is a constant, then $\ell = O(\log N)$ and $G_\ell$ has $poly(N)$ vertices. Our first lemma shows that $\mathcal{T}$ is capable of turning an input graph $G$ into an expander $G_\ell$ (as long as $H$ is in itself an expander).

**Lemma 3.2.** *Let $G$ and $H$ be the inputs of $\mathcal{T}$ as in Definition* 3.1. *If $\lambda(H) \leq 1/2$ and $G$ is connected and non-bipartite then $\lambda(\mathcal{T}(G, H)) \leq 1/2$.*

*Proof.* Since $G = G_0$ is connected and non-bipartite we have by Lemma 2.5 that $\lambda(G_0) \leq 1 - 1/DN^2$. By the choice of $\ell$ it is therefore enough to prove that for every $i > 0$, it holds that $\lambda(G_i) \leq \max\{\lambda(G_{i-1})^2, 1/2\}$. Denote $\lambda = \lambda(G_{i-1})$. Since $\lambda(H) \leq 1/2$, we have by Corollary 2.10 that $\lambda(G_{i-1} \textcircled{z} H) \leq 1 - 3/8(1 - \lambda) < 1 - 1/3(1 - \lambda)$. By the definition of $G_i$ and by Proposition 2.7 we have that $\lambda(G_i) < [1 - 1/3(1 - \lambda)]^8$. We now consider two cases. First, if $\lambda < 1/2$ then $\lambda(G_i) < (5/6)^8 < 1/2$. Otherwise, elementary calculation shows that $[1 - 1/3(1 - \lambda)]^4 \leq \lambda$ and therefore $\lambda(G_i) < \lambda^2$. The lemma follows. □

As we are working our way to solving st-connectivity, rather than solving connectivity (the problem of deciding if the input graph is connected or not), our transformation should be meaningful even for graphs that are not connected (as even in this case the two input vertices $s$ and $t$ may still be in the same connected component). For that, we will argue that $\mathcal{T}$ operates separately on each connected component of $G$. The reason is that $\mathcal{T}$ is composed of two operations (the zig-zag product and powering), that also operate separately on each connected component. We will need some additional notation: For any graph $G$ and subset of its vertices $S$, denote by $G|_S$ the subgraph of $G$ induced by $S$ (i.e., the graph on $S$ which contains all of the edges in $G$ between vertices in $S$). A set $S$ is a connected component of $G$ if $G|_S$ is connected and the set $S$ is disconnected from the rest of $G$ (i.e., there are no edges in $G$ between vertices in $S$ and vertices outside of $S$).

**Lemma 3.3.** *Let $G$ and $H$ be the inputs of $\mathcal{T}$ as in Definition* 3.1. *If $S \subseteq [N]$ is a connected component of $G$ then*

$$\mathcal{T}(G|_S, H) = \mathcal{T}(G, H)|_{S \times ([D^{16}])^\ell}.$$

*Proof.* We will only rely on $S$ being disconnected from the rest of $G$, and will prove inductively that $\mathcal{T}_i(G|_S, H) = \mathcal{T}_i(G, H)|_{S\times([D^{16}])^i}$. Note that for $i > 0$ this directly implies that $S \times ([D^{16}])^i$ is disconnected from the rest of $\mathcal{T}_i(G, H)$ (since both $\mathcal{T}_i(G|_S, H)$ and $\mathcal{T}_i(G, H)$ are $D^{16}$-regular, and thus all of the $D^{16}$ edges incident to a vertex in $S \times ([D^{16}])^i$ reside inside $\mathcal{T}_i(G, H)|_{S\times([D^{16}])^i}$). The base case $i = 0$ is trivial, and here too $S \times ([D^{16}])^i = S$ is disconnected from the rest of $\mathcal{T}_i(G, H) = G$, by assumption.

Assume by induction that $\mathcal{T}_i(G|_S, H) = \mathcal{T}_i(G, H)|_{S\times([D^{16}])^i}$. Set $G_i = \mathcal{T}_i(G, H)$ and $S_i = S \times ([D^{16}])^i$ (and recall that $S_i$ is disconnected from the rest of $G_i$). Then, by the definition of the zig-zag product, $S_i \times [D^{16}]$ is disconnected from the rest of $G_i ⓩ H$ and the edges incident to $S_i \times [D^{16}]$ in $G_i ⓩ H$ are exactly as in $G_i|_{S_i\times[D^{16}]} ⓩ H$. By the definition of powering we now have that $S_i \times [D^{16}]$ is disconnected from the rest of $(G_i ⓩ H)^8$ and the edges incident to $S_i \times [D^{16}]$ in $(G_i ⓩ H)^8$ are exactly as in $(G_i|_{S_i\times[D^{16}]} ⓩ H)^8$. This proves the induction hypothesis for $i + 1$ and completes the proof. □

Finally, we need to argue that $\mathcal{T}$ is a log-space transformation (when $D$ is a constant). The reason is that the evaluation of the rotation map $\mathrm{Rot}_{G_{i+1}}$ of each graph $G_{i+1}$ in the definition of $\mathcal{T}$ requires just a constant additional amount of memory over the evaluation of $\mathrm{Rot}_{G_i}$. Simply, the evaluation of $\mathrm{Rot}_{G_{i+1}}$ is composed of a constant number of operations, where each operation is either an evaluation of $\mathrm{Rot}_{G_i}$ or it requires constant amount of memory (and the same memory can be used for each one of these operations). So the additional memory needed for evaluating $\mathrm{Rot}_{G_{i+1}}$ is essentially a constant size counter (keeping track of which operation we are currently performing).

**Lemma 3.4.** *For every constant $D$ the transformation $\mathcal{T}$ of Definition* 3.1 *can be computed in space $O(\log N)$ on inputs $G$ and $H$, where $G$ is a $D^{16}$-regular graph on $[N]$ and $H$ is a $D$-regular graph on $[D^{16}]$.*

*Proof.* We describe an algorithm $\mathcal{A}$ that on inputs $G$ and $H$ computes the rotation map $\mathrm{Rot}_{G_\ell}$ of $G_\ell = \mathcal{T}(G, H)$. Namely, given $G$ and $H$ (written on the read-only input tape), it enumerates all values $(\bar{v}, \bar{a})$ in the domain of $\mathrm{Rot}_{G_\ell}$ and outputs $[(\bar{v}, \bar{a}), \mathrm{Rot}_{G_\ell}(\bar{v}, \bar{a})]$. Recall that a value $(\bar{v}, \bar{a})$ in the domain of $\mathrm{Rot}_{G_\ell}$ consists of $\bar{v} \in [N] \times ([D^{16}])^\ell$ which is the name of a $G_\ell$ vertex, and $\bar{a} \in [D^{16}]$, which is the label of a $G_\ell$ edge. Since $\ell = O(\log N)$ and $D$ is a constant, the length of each value $(\bar{v}, \bar{a})$ is $O(\log N)$ and therefore enumerating all of these values can be done in space $O(\log N)$. It remains to show that for any particular value $(\bar{v}, \bar{a})$, evaluating $\mathrm{Rot}_{G_\ell}(\bar{v}, \bar{a})$ can also be done in the required space.

The algorithm $\mathcal{A}$ will first allocate the following variables: $v$ which will take value in $[N]$ (specifying a vertex of $G$), and $\ell + 1$ variables $a_0, a_1, \ldots, a_\ell$ each taking value in $[D^{16}]$ (and each specifying a vertex name of $H$; In addition, $a_0$ may specify an edge label of $G$). It is sometimes convenient to view each one of $a_1, \ldots, a_\ell$ as specifying a sequence of 16 edge labels of $H$. In this case we denote $a_i = k_{i,1} \ldots k_{i,16}$. Now, $\mathcal{A}$

will copy the value $(\bar{v}, \bar{a})$ into the above mentioned variables: $\bar{v}$ into $v, a_0, \ldots, a_{\ell-1}$ and $\bar{a}$ into $a_\ell$. Throughout the execution of $\mathcal{A}$, the values of these variables will slowly evolve such that when $\mathcal{A}$ finishes (for this particular $(\bar{v}, \bar{a})$), *the same variables* will contain the desired output $\mathrm{Rot}_{G_\ell}(\bar{v}, \bar{a})$ (which is of the same range as the input $(\bar{v}, \bar{a})$).

We describe the operation of $\mathcal{A}$ in a recursive manner that closely follows the definition of $\mathcal{T}$. Particularly, at each level of the recursion, $\mathcal{A}$ will evaluate $\mathrm{Rot}_{G_i}$ for some $i$ on the appropriate prefix $v, a_0, \ldots, a_i$ of the variables defined above. For the base case $i = 0$, $\mathrm{Rot}_{G_0} = \mathrm{Rot}_G$ is written on the input tape, and can therefore be evaluated in space $O(\log N)$ by simply searching the input tape for the desired entry. For larger $i$, the evaluation of $\mathrm{Rot}_{G_i}$ is as follows:

**For $j = 1$ to 16:**

- Set $a_{i-1}, k_{i,j} \leftarrow \mathrm{Rot}_H(a_{i-1}, k_{i,j})$.
- If $j$ is odd, recursively set $v, a_0 \ldots a_{i-1} \leftarrow \mathrm{Rot}_{G_{i-1}}((v, a_0 \ldots a_{i-2}), a_{i-1})$.
- If $j = 16$, reverse the order of the individual labels in $a_i$: Set $k_{i,1}, \ldots, k_{i,16} \leftarrow k_{i,16}, \ldots, k_{i,1}$.

The correctness of $\mathcal{A}$ immediately follows from the definition of $\mathcal{T}$ and from the operations of which it consists (powering and the zig-zag product). Essentially, going over the operations (in the first two bullets) for any two consecutive values of $j$ corresponds to one step on $(G_{i-1} ⓩ H)$. Repeating eight times implies a path of length eight on $(G_{i-1} ⓩ H)$, or alternatively one step on $(G_{i-1} ⓩ H)^8$. The third bullet reverses the order of labels to fit the definition of zig-zag and powering.

We therefore concentrate on the space complexity of $\mathcal{A}$. Note that each node of the recursion tree performs a constant number of operations and makes a constant number of recursive calls. In addition the depth of the recursion is $\ell + 1 = O(\log N)$. Therefore, maintaining the recursion can be done in space $O(\log N)$. Furthermore, each one of the basic operations (evaluating $\mathrm{Rot}_G$, evaluating $\mathrm{Rot}_H$, and reversing the order of labels in the last step) can be performed in space $O(\log N)$. Finally, the only memory that needs to be kept after a basic operation is performed, is the memory holding the variables $v, a_0, \ldots, a_\ell$ (that are shared by all of these operations), and the memory for maintaining the recursion. We therefore conclude that the space complexity of $\mathcal{A}$ is $O(\log N)$ which completes the proof. □

## 4. A log-space algorithm for USTCON

This section puts together the tools developed above into a deterministic log-space algorithm that decides undirected st-connectivity. As will be discussed in Section 5, the algorithm can also output a path from $s$ to $t$ if such a path exists.

**Theorem 4.1.** USTCON $\in$ L.

As undirected USTCON is complete for SL [25], Theorem 4.1 can be rephrased as follows.

**Theorem 4.2.** SL = L.

*Proof of Theorem* 4.1. We give an algorithm $\mathcal{A}$ that gets as input a graph $G$ over the set of vertices $[N]$, and two vertices $s$ and $t$ in $[N]$. For concreteness, we assume that the graph is given via the adjacency matrix representation. $\mathcal{A}$ will answer 'connected' if and only if there exists a path in $G$ between $s$ and $t$ (i.e., $s$ and $t$ are in the same connected component). Furthermore, $G$ will use space which is logarithmic in its input size.

The algorithm $\mathcal{A}$ will need to evaluate the rotation map of a $((D_e)^{16}, D_e, 1/2)$-graph $H$, where $D_e$ is some constant. By Proposition 2.4, there exists such a graph and therefore $\mathcal{A}$ can obtain it by exhaustive search using constant amount of memory (a more efficient alternative is of course to obtain $H$ by any of the explicit constructions of expanders mentioned in Section 2.2).

Let $\mathcal{T}$ be the transformation given by Definition 3.1. We would like to apply $\mathcal{T}$ to $G$ and $H$ in order to obtain a graph where each connected component is an expander. For such graphs, $st$-connectivity can be solved in logarithmic space by Proposition 2.3. However, we will first need to preprocess $G$ in order to get a new graph $G_{\text{reg}}$ such that $(G_{\text{reg}}, H)$ is a correct input to $\mathcal{T}$. In particular, we need $G_{\text{reg}}$ to be a $D_e^{16}$-regular graph given by its rotation map. There are various ways of transforming $G$ to $G_{\text{reg}}$. The one given here was selected for its simplicity even though it is not the most efficient one possible (in terms of the size of $G_{\text{reg}}$). Essentially, we replace every vertex of $G$ with a cycle of length $N$ and each of the vertices $(v, w)$, where there is an edge between $v$ and $w$ in $G$, is also connected to $(w, v)$ (the rest of the edges are self loops). The rotation map $\text{Rot}_{G_{\text{reg}}}: ([N] \times [N]) \times [D_e^{16}] \mapsto ([N] \times [N]) \times [D_e^{16}]$ of $G_{\text{reg}}$ is formally defined as follows:

- $\text{Rot}_{G_{\text{reg}}}((v, w), 1) = ((v, w'), 2)$, where $w' = w + 1$ if $w < N$ and $w' = 1$ otherwise.
- $\text{Rot}_{G_{\text{reg}}}((v, w), 2) = ((v, w'), 1)$, where $w' = w - 1$ if $w > 1$ and $w' = N$ otherwise.
- In case there is an edge between $v$ and $w$ in $G$ then $\text{Rot}_{G_{\text{reg}}}((v, w), 3) = ((w, v), 3)$. Otherwise, $\text{Rot}_{G_{\text{reg}}}((v, w), 3) = ((v, w), 3)$.
- For $i > 3$, $\text{Rot}_{G_{\text{reg}}}((v, w), i) = ((v, w), i)$.

The transformation from $G$ (given by its adjacency matrix) to $G_{\text{reg}}$ (given by its rotation map) is clearly computable in logarithmic space. Furthermore, $G_{\text{reg}}$ is $D_e^{16}$-regular by definition and all its connected components are non-bipartite (as every vertex in $G_{\text{reg}}$ has self loops). Finally, for every connected component $S \subseteq [N]$ of $G$ we have that $S \times [N]$ is a connected component in $G_{\text{reg}}$. To see that, we first note that for every vertex $v \in [N]$ the set of vertices $v \times [N]$ is in the same connected

component of $G_{\text{reg}}$ (as this set is connected by a cycle). Furthermore, there is an edge in $G_{\text{reg}}$ between some vertex in $v \times [N]$ and some vertex in $w \times [N]$ if and only if $v$ and $w$ are connected by an edge in $G$ (the only possible edge that can connect these subsets is an edge between $(v, w)$ and $(w, v)$ which only exists in $G_{\text{reg}}$ if there is an edge between $v$ and $w$ in $G$).

Now define $G_{\text{exp}} = \mathcal{T}(G_{\text{reg}}, H)$, and $\ell = O(\log N)$ is the corresponding value as in Definition 3.1. Let $S$ be the connected component of $G$, such that $s \in S$. By the arguments above, $S \times [N]$ is a connected component of $G_{\text{reg}}$, and $G_{\text{reg}}|_{S\times[N]}$ is non-bipartite. By Lemma 3.3, $S \times [N] \times ([D^{16}])^{\ell}$ is a connected component of $G_{\text{exp}}$ (as both $G_{\text{exp}}$ and $G_{\text{exp}}|_{S\times[N]\times([D^{16}])^{\ell}}$ are $D_{\text{e}}^{16}$-regular). By Lemma 3.2 and Lemma 3.3, we have that $\lambda(G_{\text{exp}}|_{S\times[N]\times([D^{16}])^{\ell}}) \leq 1/2$.

Let $\mathcal{A}'$ be the algorithm guaranteed by Proposition 2.3 (which decides undirected st-connectivity correctly in graphs where the connected component of the starting vertex is an expanders). The algorithm $\mathcal{A}$ will now invoke $\mathcal{A}'$, on the graph $G_{\text{exp}}$ and the vertices $s' = (s, 1^{\ell+1})$ and $t' = (t, 1^{\ell+1})$. If $\mathcal{A}'$ outputs that $s'$ and $t'$ are connected in $G_{\text{exp}}$ then $\mathcal{A}$ will output that $s$ and $t$ are connected in $G$. Otherwise, $\mathcal{A}$ will output that $s$ and $t$ are not connected.

The algorithm $\mathcal{A}$ is log-space since it is composed of a constant number of log-space procedures: (1) The transformation from $G$ to $G_{\text{reg}}$. (2) The transformation from $G_{\text{reg}}$ to $G_{\text{exp}}$, which is log-space by Lemma 3.4. (3) The algorithm $\mathcal{A}'$ which is log-space by Proposition 2.3. Correctness of $\mathcal{A}$ is argued as follows. First, $s'$ and $t'$ are connected in $G_{\text{exp}}$ if and only if $s$ and $t$ are connected in $G$ (since $S \times [N] \times ([D^{16}])^{\ell}$ is a connected component of $G_{\text{exp}}$, where $S$ is the connected component of $G$ that contains $s$). The correctness of $\mathcal{A}$ now follows since Proposition 2.3 implies that $\mathcal{A}'$ will output 'connected' if and only if $s'$ and $t'$ are indeed connected in $G_{\text{exp}}$ (as $\lambda(G_{\text{exp}}|_{S\times[N]\times([D^{16}])^{\ell}}) \leq 1/2$). □

## 5. Universal traversal and exploration sequences

In this section, we look closer into our USTCON algorithm and conclude that it also solves the corresponding search problem (i.e., finding the path from $s$ to $t$ if such a path exist). In addition, it implies efficiently-constructible universal-traversal sequences for graphs with restricted labelling, and universal exploration sequences for general graphs. The sort of restriction we pose on the labelling of graphs is a strengthening of the "consistent labelling" used in [18]. In a subsequent work [41], our restriction is relaxed to that of [18].

We start by analyzing $\mathcal{T}$, the main transformation of the algorithm, given by Definition 3.1. We show that every edge in $\mathcal{T}(G, H)$ translates to a path in $G$ between the appropriate vertices, and that this path is log-space constructible (as this path is indeed computed during the log-space evaluation of $\mathcal{T}$). Looking ahead to the universal-traversal sequences, we note that if we restrict the labelling of $G$, then the labels of edges, traversed along this path, are independent of $G$.

**Definition 5.1.** Let $\pi$ be a permutation over $[D]$ and $\mathrm{Rot}_G$ the rotation map of a $D$-regular graph $G$. Then $\mathrm{Rot}_G$ is $\pi$-consistent if for every $v, i, w$ and $j$ such that $\mathrm{Rot}_G(v, i) = (w, j)$, it holds that $j = \pi(i)$. In such a case we may also say that the labelling of $G$ is $\pi$-consistent.

An example of a $\pi$-consistent labelling is symmetric labelling where $\pi$ is simply the identity. Namely, every edge is labelled in the same way from both its end points. However, other kinds of $\pi$-consistent labellings come up naturally. An example for that is the labelling of $G_{\mathrm{reg}}$ in the proof of Theorem 4.1. We can now state the appropriate technical lemma regarding the transformation $\mathcal{T}$.

**Lemma 5.2.** *Let $D$ be some constant. Let $G$ be a $D^{16}$-regular graph on $[N]$ and let $H$ be a $D$-regular graph on $[D^{16}]$, both given by their rotation maps. Let $G_\ell = \mathcal{T}(G, H)$, where $\mathcal{T}$ and $\ell$ are given by Definition* 3.1.

*There exists a log-space algorithm such that given $\mathrm{Rot}_G$, $\mathrm{Rot}_H$ and $(\bar{v}, \bar{a})$ in the domain of $\mathrm{Rot}_{G_\ell}$, it outputs a sequence of labels in $[D^{16}]$ with the following property: If the first element of $\bar{v}$ is a vertex $u \in [N]$ and the first element of $\mathrm{Rot}_{G_\ell}(\bar{v}, \bar{a})$ is a vertex $w \in [N]$, then the walk on $G$ from $u$ using the labels that the algorithm outputs leads to $w$.*

*Furthermore, for every fixed permutation $\pi$ on $[D^{16}]$, if the labelling of $G$ is $\pi$-consistent, the log-space algorithm can evaluate the sequence of labels without access to $\mathrm{Rot}_G$.*

*Proof.* Consider the log-space algorithm $\mathcal{A}$ in the proof of Theorem 3.4, as it evaluates $\mathrm{Rot}_{G_\ell}(\bar{v}, \bar{a})$. We enhance it a bit, to define an algorithm $\mathcal{A}'$ as claimed by the lemma. Consider in particular the two variables $v$ and $a_0$ used by $\mathcal{A}$. To begin with, $v$ will be initialized to the value $u$ (the first element of $\bar{v}$). At the end, $v$ will contain the value $w$. Throughout the run of $\mathcal{A}$, the variable $v$ is only updated by the rule $v, a_0 \leftarrow \mathrm{Rot}_G(v, a_0)$ (used at the bottom of the recursion). Therefore, all that $\mathcal{A}'$ needs to do is to output the value of $a_0$ just before each time $\mathcal{A}$ updates $v$.

Regarding the second part of the lemma. We note that the value of $a_0$ is only influenced by $\mathrm{Rot}_G$, through the evaluations $v, a_0 \leftarrow \mathrm{Rot}_G(v, a_0)$. If $G$ is $\pi$-consistent, then $\mathcal{A}'$ can completely ignore the variable $v$ and the rotation map of $G$. To simulate $\mathcal{A}$, it is sufficient that whenever $\mathcal{A}$ evaluates $v, a_0 \leftarrow \mathrm{Rot}_G(v, a_0)$, then $\mathcal{A}'$ will evaluate $a_0 \leftarrow \pi(a_0)$. □

Using Lemma 5.2, it is not hard to obtain the algorithm that finds paths in undirected graphs.

**Theorem 5.3.** *There exists a log-space algorithm that gets as input a graph $G$ over the set of vertices $[N]$, and two vertices $s$ and $t$ in $[N]$, and outputs a path from $s$ to $t$ if such a path exists (otherwise it outputs 'not connected').*

*Proof.* Consider the algorithm $\mathcal{A}$ from the proof of Theorem 4.1. We revise it to an algorithm $\mathcal{A}'$ as required by the theorem. First, we note that it is enough for $\mathcal{A}'$ to

output a path from $(s, 1)$ to $(t, 1)$ in $G_{\text{reg}}$ if such a path exists, as it is easy to transform (in log-space) such a path to a path from $s$ to $t$ in $G$ (and the existence of the two paths is equivalent).

Next we note that $\mathcal{A}$ enumerates all logarithmically-long paths from $s' = (s, 1^{\ell+1})$ in $G_{\text{exp}}$. If it does not find a path that visits $t' = (t, 1^{\ell+1})$, it concludes that $s$ and $t$ are not connected in $G$. Therefore, in such a case, $\mathcal{A}'$ can output 'not connected'. Otherwise $\mathcal{A}$ found a short path from $s'$ to $t'$. Apply the algorithm guaranteed by Lemma 5.2 on each edge on the path from $s'$ to $t'$. Each time the algorithm outputs a sequence of edge-labels in $G_{\text{reg}}$. Let $\vec{a}$ be the concatenation of these sequences. It follows from Lemma 5.2 that the path in $G_{\text{reg}}$ starting from $(s, 1)$ and following the edges according to the labels in $\vec{a}$ leads to $(t, 1)$. The theorem now follows. □

To give our result regarding universal-traversal sequences, we need some notation. Let $\vec{a} = \{a_1, \ldots, a_m\}$ be a sequence of values in $[D]$ (these are interpreted as edge labels). $\vec{a}$ is an $(N, D)$-universal traversal sequence, if for every connected $D$-regular, labelled graph $G$ on $N$ vertices, and every start vertex $s \in [N]$, the walk that starts at $s$ and follows the edges labelled $a_1, \ldots, a_m$, visits every vertex in the graph. For a permutation $\pi$ over $[D]$, we say that $\vec{a}$ is an $(N, D)$ $\pi$-universal traversal sequence, if the above property holds for every connected $D$-regular graph on $N$ vertices, *that has a $\pi$-consistent labelling*, (rather than for all such graphs).

**Theorem 5.4.** *There exists a log-space algorithm that takes as input $1^N$ and a permutation $\pi$ over $[D]$ and outputs an $(N, D)$ $\pi$-universal traversal sequence.*

*Proof.* First we argue that it is enough to construct an $(N \cdot D, D_{\text{e}}^{16})$ $\pi'$-universal sequence for the following simple permutation: $\pi'(1) = 2$, $\pi'(2) = 1$ and for every $i > 2$ $\pi'(i) = i$. Furthermore, all we need is that the sequence will traverse non-bipartite graphs. Consider a (connected) $D$-regular graph $G$ on $N$ vertices that has a $\pi$-consistent labelling. This graph can be transformed into a $D_{\text{e}}^{16}$-regular (connected and non-bipartite) graph $G'$ on $N \cdot D$ vertices that has a $\pi'$-consistent labelling. Each vertex $v \in N$ is transformed into a cycle over $D$ vertices $(v, 1), \ldots, (v, D)$, the edges of the cycle are labelled 1 and 2 (just as in the definition of $G_{\text{reg}}$ in the proof of Theorem 4.1). The edge labelled 3 going out of $(v, i)$ will lead to $\text{Rot}_G(v, i)$ (and will be labelled 3 from that end as well). All other edges are self loops.

Assume that a sequence of labels $a_1, \ldots, a_m$, visits every vertex of $G'$ starting from every vertex $(v, 1)$ (this is even less general than what we obtain). We can translate this (in log space) into a sequence of labels $b_1, \ldots, b_{m'}$ that traverses $G$ from every vertex $v$. To do that, we simulate the walk on $G'$ from an arbitrary vertex $(v, 1)$. As $v$ is unknown and our simulation does not rely on $G$, it will only know at each point the value $b$ such that the walk at this point visits some vertex $(w, b)$ of $G'$ (where $w$ is unknown). First $b$ is set to 1. Then, during the simulation, labels $a_i > 3$ can be ignored (as they are self loops). Given labels 1 and 2, $b$ can easily be updated (these are edges on the cycle). Finally, when encountering $a_i = 3$ the walk moves from a vertex $(w, b)$ to a vertex $(w', \pi(b))$ (as the labelling of $G$ is $\pi$-consistent), and so it is

easy to update the value of $b$ (given access to $\pi$). The projection of the walk on $G$ is exactly the edges labelled 3 that are taken by the walk on $G'$. Therefore, to transform the sequence of $a_i$'s to the sequence of $b_i$'s we can simply output (throughout the simulation) the current value of $b$, whenever we encounter a label $a_i = 3$.

Now we consider a $D_e^{16}$-regular (connected and non-bipartite) graph $G'$ on $N \cdot D$ vertices that has a $\pi'$-consistent labelling. Let $H$ be a $((D_e)^{16}, D_e, 1/2)$-graph. Finally let $G_\ell = \mathcal{T}(G, H)$, where $\mathcal{T}$ and $\ell$ are given by Definition 3.1. By Lemma 3.2, $\lambda(G_\ell) \leq 1/2$ and therefore its diameter is logarithmic. Therefore, for every two vertices $v$ and $u$ of $G'$ one of the polynomially many sequences of labels (of the appropriate logarithmic length) will visit $(u, 1^\ell)$, starting at $(v, 1^\ell)$. Let $B$ be the set of all these sequences of labels. Lemma 5.2 gives a way to translate in log-space each one of the sequences in $B$ into a corresponding sequence of edge-labels of $G'$. Let $B'$ be the set of translated sequences. By Lemma 5.2 and the above argument, for every two vertices $v$ and $u$ of $G'$ one of the sequences in $B'$ will lead a walk in $G'$ that starts in $v$ through the vertex $u$. We should also note that given a sequence $\vec{a} = a_1, \ldots, a_m$ that leads from a vertex $v$ to a vertex $u$, we have that the sequence $\pi'^{-1}(a_m), \ldots, \pi'^{-1}(a_1)$ leads from $u$ to $v$ (this operation simply reverses the walk). We refer to this latter sequence as the reverse of $\vec{a}$. Finally, we can define a sequence that traverses all of the vertices of $G'$ regardless of the starting vertex. Simply, we concatenate for each sequences in $B'$ its reversed sequence and concatenate all of these sequences one after the other. By the arguments above, for every vertex $v$, the sequence we obtain will visit $v$ after every pair of a sequence and its reversed sequence. Furthermore, for every vertex $u$, one of these sequences will lead to $u$. As the log-space construction of this sequence ignores the graph $G'$ (and only relies on $\pi'$), we obtained the desired $(N \cdot D, D_e^{16})$ $\pi'$-universal sequence for non-bipartite graphs. The theorem follows. □

In an $(N, D)$-universal *exploration* sequence, the sequence of labels is interpreted as offsets rather than absolute labels. This means that if we entered a vertex $v$ on an edge labelled $a$ (from $v$'s view point), and we are reading the label $b$, then we will leave $v$ on the edge labelled $a + b$ (or $a + b - D$ if $a + b > D$). In fact this notion can apply to graphs that are not-regular (it then makes sense to allow negative elements in the sequence). Universal-exploration sequences have more flexibility than universal-traversal sequences. For example, it is not clear how to transform a universal-traversal sequence for degree-3 graphs to one for higher-degree graphs. This is easy for universal-exploration sequences (and seems desirable as USTCON can easily be reduced to USTCON for regular-graphs of any degree larger than 2). Koucky [24] showed how to transform a universal-traversal sequence to a universal-exploration sequence. His transformation (which is essentially the same as the one from $G$ to $G'$ in the proof of Theorem 5.4), only needs the universal-sequence to work for graphs with $\pi$-consistent labelling for some simple permutation $\pi$. We can therefore conclude from Theorem 5.4 a log-space construction for general universal-exploration sequences.

**Corollary 5.5.** *There exists a log-space algorithm that takes as input* $(1^N, 1^D)$ *and produces an* $(N, D)$*-universal exploration sequence.*

## 6. Discussion and further research

We start by comparing the techniques of this paper with some previous ones, with the goal of shading some light on the source of our improvements. We continue by discussing some open problems and the results of a subsequent work.

**Comparison with previous techniques** The USTCON algorithms of [47], [34], [9] also operate by transforming, in phases, the input graph into a more accommodating one. In each one of these algorithms, each phase "charges" logarithmic amount to the space complexity of the algorithm. The improvement in the space complexity is directly correlated to reducing the number of phases needed for the transformation. With this approach, the only way to obtain a log-space algorithm is to reduce the number of phases to a constant. We deviate from this direction, as we use a logarithmic number of phases (just as in Savitch's algorithm), to gradually improve the connectivity of the input graph. The space efficiency of our algorithm stems from each transformation being significantly less costly in space.

The parameter being improved by [34], [9], is the size of the graph (each transformation shrinks the graph by collapsing it to a "representative" subset of the vertices). In contrast, our transformation will in fact expand the graph by a polynomial factor (as each phase, enlarges our graph by a constant factor). The parameter Savitch's transformation improves is the diameter of the graph, which is much closer to the parameter we improve (the expansion). In fact, each phase of Savitch's algorithm can be described very similarly to our algorithm. Each one of these phases consists of squaring the graph and then removing parallel edges (which may reduce the degree). Although all that is needed is indeed that the *diameter* of the resulting graph will be small, our analysis relies on bounding the *expansion* of intermediate graphs – a stronger notion of connectivity than the diameter. This allows our transformation to preserve constant degree of the graph (rather than linear degree in Savitch's algorithm), which is crucial for our analysis of the space complexity.

It also seems instructive to compare with the combinatorial construction of expander graphs of [43]. There, an arbitrarily large expander graphs was constructed, starting with a constant size expander. This small expander is made larger and larger, while its degree is kept constant via the zig-zag or the replacement product. Our main transformation shows how to turn *any* connected graph (which is already large) into an expander. This means that the above mentioned products need to be applied when one of the graphs is an extremely weak expander (whereas in [43] both graphs were fairly good expanders). Very fortunately, both products work quite well in this unusual setting of parameters.

**Further Research** There are many open problems and new research directions brought up by this work, we discuss just a few of those. A very natural question is whether the techniques of this paper can be used towards a proof of RL = L. While progress in the context of RL does not seem immediate (as the case of symmetric computations does seem easier), we feel that it is still quite plausible. We also feel that this paper should give an opportunity to reevaluate the common conjecture that Savitch's algorithm is optimal for STCON. While this conjecture may very well be correct, we feel that there is not enough evidence supporting it. Another open problem is to come up with full-fledged, efficiently-constructible, universal-traversal sequences. Interestingly, it seems that this problem shares some of the obstacles that one encounters when trying to generalize the USTCON algorithm to solving RL (this is formalized to some extent in the results of [41] mentioned below).

In a subsequent work, Reingold, Trevisan and Vadhan [41], make some progress on extending our techniques to dealing with the general RL case, obtaining the following results:

1. They exhibit a new complete problem for RL: STCON restricted to directed graphs for which the random walk is promised to have polynomial mixing time.
2. Generalizing our techniques, they present a deterministic, log-space algorithm that given a *regular* directed graph $G$ (i.e., a directed graph where all in-degrees and out-degrees are equal) and two vertices $s$ and $t$, finds a path between $s$ and $t$ if one exists.
3. Using the same techniques as in Item 2, they give a "pseudorandom generator" for random walks on "consistently labelled" regular directed graphs. Roughly speaking, given a random seed of logarithmic length, the generator constructs, in log-space, a "short" pseudorandom walk that ends at an almost-uniformly distributed vertex when taken in any consistently labelled regular directed graph.
4. They prove that if their pseudorandom generator from Item 3 could be generalized to all regular directed graphs (instead of just consistently labelled ones), then their complete problem from Item 1 can be solved in log-space and hence RL = L.

Finally, we have made no attempt to optimize our algorithm in terms of running time (or the constant in the space complexity). Major improvements in efficiency can come about by better analysis of the zig-zag and replacement products. These may also determine which one of these products yields a more efficient algorithm. In a subsequent work Rozenman and Vadhan [44] give a log-space algorithm for USTCON. Their algorithm makes substantial progress in terms of reducing the running time of the algorithm (compared to ours). Their key technical tool is a new operation they introduce and name "derandomized squaring". This operation reduces the second eigenvalue of a graph "similarly" to standard squaring but increases the degree much more moderately. Very loosely, this operation can replace in our algorithm the combination of (standard) powering and zig-zag product. Their analysis for

the new operation is tight (unlike the analysis we currently know for the zig-zag and replacement products) and it is simple and very appealing.[6]

**Acknowledgments.** This work came about during a delightful visit to UC Berkeley. I am most grateful to Irit Dinur and Luca Trevisan for many hours of stimulating discussions on closely related topics and for creating the most conducive research environment possible for me. I would like to thank Moni Naor, Ran Raz, Salil Vadhan and Avi Wigderson for many discussions that helped me form my intuitions on the derandomization of space bounded computations. Among other contributions, I want to thank Moni for steering me towards this topic early on during my PhD studies, and to thank Ran, Salil and Avi for intuitions formed during our joint work on [38], [42].

## References

[1] Ajtai, Miklós, Komlós, János, and Szemerédi, E., Deterministic simulation in LOGSPACE. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1987, 132–140.

[2] Aleliunas, Romas, Karp, Richard M., Lipton, Richard J., Lovász, László, and Rackoff, Charles, Random walks, universal traversal sequences, and the complexity of maze problems. In *Proceedings of the 20th Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1979, 218–223.

[3] Alon, Noga, Eigenvalues and expanders. *Combinatorica* **6** (2) (1986), 83–96.

[4] Alon, Noga, Galil, Zvi, and Milman, Vitali D., Better expanders and superconcentrators. *J. Algorithms* **8** (3) (1987), 337–347.

[5] Alon, Noga, and Milman, Vitali D., $\lambda_1$, isoperimetric inequalities for graphs, and superconcentrators. *J. Combin. Theory Ser. B* **38** (1) (1985), 73–88.

[6] Alon, Noga, and Roichman, Yuval, Random Cayley graphs and expanders. *Random Structures Algorithms* **5** (2) (1994), 271–284.

[7] Alon, Noga, and Sudakov, Benny, Bipartite subgraphs and the smallest eigenvalue. *Combin. Probab. Comput.* **9** (1) (2000), 1–12.

[8] Àlvarez, Carme, and Greenlaw, Raymond, A compendium of problems complete for symmetric logarithmic space. *Comput. Complexity* **9** (2000), 123–145.

[9] Armoni, Roy, Ta-Shma, Amnon, Wigderson, Avi, and Zhou, Shiyu, An $o(log(n)^{4/3})$ space algorithm for $(s, t)$ connectivity in undirected graphs. *J. ACM* **47** (2), (2000), 294–311.

[10] Babai, László, Nisan, Noam, and Szegedy, Márió, Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs. *J. Comput. System Sci.* **45** (1992), 204–232.

[11] Ben-Asher, Y., Lange, K., Peleg, D., and Schuster, A., The complexity of reconfiguring network models. *Inform. and Comput.* **21** (1) (1995), 41–58.

[6] In essence, their analysis implies that an expander can be viewed as a convex combination of a complete graph (the ultimate expander) and an error term. This insightful observation can be exploited to simplify the analysis of the zig-zag and replacement products as well (but the obtained analysis is still not tight).

[12] Borodin, Allan, Cook, Stephen A., Dymond, Patrick W., Ruzzo, Walter L., and Tompa, Martin, Two applications of inductive counting for complementation problems. *SIAM J. Comput.* **18** (3) (1989), 559–578.

[13] Broder, Andrei, and Shamir, Eli, On the second eigenvalue of random regular graphs. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1987, 286–294.

[14] Friedman, Joel, On the second eigenvalue and random walks in random $d$-regular graphs. *Combinatorica* **11** (4) (1991), 331–362.

[15] Friedman, Joel, Kahn, Jeff, and Szemerédi, Endre, On the second eigenvalue in random regular graphs. In *Proceedings of the 21th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1989, 587–598.

[16] Gabber, Ofer, and Galil, Zvi, Explicit constructions of linear-sized superconcentrators. *J. Comput. System Sci.* **22** (3) (1981), 407–420.

[17] Goldreich, Oded, and Wigderson, Avi, Derandomization that is rarely wrong from short advice that is typically good. In *RANDOM 2002*, Lecture Notes in Comput. Sci. 2483, Springer-Verlag, Berlin 2002, 209–223.

[18] Hoory, Shlomo, and Wigderson, Avi, Universal traversal sequences for expander graphs. *Inform. Process. Lett.* **46** (2) (1993), 67–69.

[19] Impagliazzo, Russell, Nisan, Noam, and Wigderson, Avi, Pseudorandomness for network algorithms. In *Proceedings of the 26th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 1994, 356–364.

[20] Jimbo, Shuji, and Maruoka, Akira, Expanders obtained from affine transformations. *Combinatorica* **7** (4) (1987), 343–355.

[21] Karchmer, Mauricio, and Wigderson, Avi, On span programs. In *Proceedings of the 8th Structures in Complexity Conference*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1993, 102–111.

[22] Klivans, Adam, and van Melkebeek, Dieter, Graph nonisomorphism has subexponential size proofs unless the polynomial-time hierarchy collapses. *SIAM J. Comput.* **31** (5) (2002), 1501–1526.

[23] Koucky, Michal, Universal traversal sequences with backtracking. *J. Comput. System Sci.* **65** (2002), 717–726.

[24] Koucky, Michal, *On traversal sequences, exploration sequences and completeness of Kolmogorov random strings*. PhD thesis, Rutgers University, 2003.

[25] Lewis, Harry R., and Papadimitriou, Christos H., Symmetric space-bounded computation. *Theoret. Comput. Sci.* **19** (1982), 161–187.

[26] Lubotzky, Alex, Phillips, Ralph, and Sarnak, Peter, Ramanujan graphs. *Combinatorica* **8** (3) (1988), 261–277.

[27] Madras, Neal, and Randall, Dana, Factoring graphs to bound mixing rates. In *Proceedings of the 37th Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1996, 194–203.

[28] Margulis, Gregory A., Explicit constructions of expanders. *Problemy Peredači Informacii* **9** (4) (1973), 71–80.

[29] Margulis, Gregory A., Explicit group-theoretic constructions of combinatorial schemes and their applications in the construction of expanders and concentrators. *Problemy Peredachi Informatsii* **24** (1) (1988), 51–60.

[30] Martin, Russell A., and Randall, Dana, Sampling adsorbing staircase walks using a new markov chain decomposition method. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2000, 492–502.

[31] Morgenstern, Moshe, Existence and explicit constructions of $q + 1$ regular Ramanujan graphs for every prime power $q$. *J. Combin. Theory. Series B* **62** (1) (1994), 44–62.

[32] Nisan, Noam, RL $\subseteq$ SC. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 1992, 619–623.

[33] Nisan, Noam, Pseudorandom generators for space-bounded computation. *Combinatorica* **12** (4) (1992), 449–461.

[34] Nisan, Noam, Szemeredi, Endre, and Wigderson, Avi, Undirected connectivity in $o(\log^{1.5} n)$ space. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1989, 24–29.

[35] Nisan, Noam, and Ta-Shma, Amnon, Symmetric logspace is closed under complement. *Chicago J. Theor. Comput. Sci.*, 1995 (electronic).

[36] Nisan, Noam, and Zuckerman, David, Randomness is linear in space. *J. Comput. System Sci.* **52** (1) (1996), 43–52.

[37] Pinsker, Mark S., On the complexity of a concentrator. In *7th Annual Teletraffic Conference*, Stockholm, 1973, 318/1–318/4.

[38] Raz, Ran, and Reingold, Omer, On recycling the randomness of the states in space bounded computation. In *Proceedings of the 31st Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 1999, 159–168.

[39] Reif, John H., Symmetric complementation. *J. ACM* **31** (2) (1984), 401–421.

[40] Reingold, Omer, Undirected st-connectivity in log-space. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 376–385.

[41] Reingold, Omer, Trevisan, Luca, and Vadhan, Salil, Pseudorandom walks in biregular graphs and the RL vs. L problem. *Electronic Colloquium on Computational Complexity* Technical Report TR05-022, 2005.

[42] Reingold, Omer, Vadhan, Salil, and Wigderson, Avi, Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2000, 3–13.

[43] Reingold, Omer, Vadhan, Salil, and Wigderson, Avi, Entropy waves, the zig-zag graph product, and new constant-degree expanders. *Ann. of Math.* **155** (1) (2002), 157–187.

[44] Rozenman, Eyal, and Vadhan, Salil, Derandomized squaring of graphs. In *Approximation, randomization and combinatorial optimization* (Berkeley, CA, 2005), ed. by Chandra Chekuri et al., Lecture Notes in Comput. Sci. 3624, Springer-Verlag, Berlin 2005, 436–447.

[45] Saks, Michael, Randomization and derandomization in space-bounded computation. In *Proceedings of the 11th Annual Conference on Structure in Complexity Theory*, IEEE Comput. Soc. Press, Los Alamitos, CA, 1996, 128–149.

[46] Saks, Michael, and Zhou, Shiyu, $\mathrm{bp_h space}(S) \subseteq \mathrm{dspace}(S^{3/2})$. *J. Comput. System Sci.* **58** (2) (1999), 376–403.

[47] Savitch, J., Relationship between nondeterministic and deterministic tape complexities. *J. Comput. System Sci.* **4** (2) (1970), 177–192.

[48] Tanner, Michael R., Explicit concentrators from generalized $n$-gons. *SIAM J. Algebraic Discrete Methods* **5** (3) (1984), 287–293.

[49] Trifonov, Vladimir, An o(log n log log n) space algorithm for undirected s,t-connectivity. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 626–633.

[50] Wigderson, Avi, The complexity of graph connectivity. In *Mathematical Foundations of Computer Science 1992* (ed. by I. M. Havel and V. Koubek), Lecture Notes in Comput. Sci. 629, Springer-Verlag, Berlin 1992, 112–132.

Department of Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel
E-mail: omer.reingold@weizmann.ac.il

# Potential functions and the inefficiency of equilibria

Tim Roughgarden*

**Abstract.** We survey one area of the emerging field of algorithmic game theory: the use of approximation measures to quantify the inefficiency of game-theoretic equilibria. Potential functions, which enable the application of optimization theory to the study of equilibria, have been a versatile and powerful tool in this area. We use potential functions to bound the inefficiency of equilibria in three diverse, natural classes of games: selfish routing networks, resource allocation games, and Shapley network design games.

**Mathematics Subject Classification (2000).** 68Q25, 68W25, 90B35, 91A65.

**Keywords.** Game theory, inefficiency of equilibria, Nash equilibrium, network design, potential functions, price of anarchy, price of stability, resource allocation, selfish routing.

## 1. Introduction

The interface between theoretical computer science and microeconomics, often called *algorithmic game theory*, has been an extremely active research area over the past few years. Recent points of contact between the two fields are diverse and include, for example, increased attention to computational complexity and approximation in combinatorial auctions (e.g. [9]); a new focus on worst-case analysis in optimal auction design (e.g. [17]); and a renewed emphasis on the computability and learnability of equilibrium concepts (e.g. [14], [18], [26]). This survey touches on just one connection between theoretical computer science and game theory: the use of approximation measures to quantify the inefficiency of game-theoretic equilibria.

**1.1. Quantifying the inefficiency of equilibria.** Even in very simple settings, selfish behavior can lead to highly inefficient outcomes [11]. A canonical example of this phenomenon is provided by the "Prisoner's Dilemma" [28], in which strategic behavior by two captured and separated prisoners inexorably draws them into the worst-possible outcome. We will see several concrete examples of the inefficiency of selfish behavior in networks later in the survey.

Must more recently, researchers have sought to *quantify* the inefficiency of selfish behavior. Koutsoupias and Papadimitriou [23] proposed a framework to systematically study this issue. The framework presupposes a strategic environment (a *game*), a definition for the outcome of selfish behavior (an *equilibrium concept*), and a real-

*Supported in part by ONR grant N00014-04-1-0725, DARPA grant W911NF-04-9-0001, and an NSF CAREER Award.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

valued, nonnegative objective function defined on the possible outcomes of the game. The *price of anarchy* [23], [26] is then defined as the ratio between the objective function value of an equilibrium and that of an optimal solution. (For the moment, we ignore the question of whether or not equilibria exist and are unique.) If the price of anarchy of a game is 1, then its equilibria are fully efficient. More generally, bounding the price of anarchy in a class of games provides a guarantee on the worst-possible inefficiency of equilibria in these games.

The price of anarchy is directly inspired by other popular notions of approximation in theoretical computer science [23]. One example is the *approximation ratio* of a heuristic for a (typically NP-hard) optimization problem, defined as the largest ratio between the objective function value of the solution produced by the heuristic and that of an optimal solution. While the approximation ratio measures the worst-case loss in solution quality due to insufficient computational effort, the price of anarchy measures the worst-case loss arising from insufficient ability (or willingness) to control and coordinate the actions of selfish individuals. Much recent research on the price of anarchy is motivated by optimization problems that naturally occur in the design and management of large networks (like the Internet), in which users act selfishly, but implementing an optimal solution is not practical.

**1.2. Potential functions.** The price of anarchy has been successfully analyzed in a diverse array of game-theoretic models (see e.g. [32], [33] and the references therein). This survey discusses three of these models, with the goal of illustrating a single mathematical tool for bounding the price of anarchy: *potential functions*. The potential function technique is by no means the only one known for bounding the inefficiency of equilibria, but (so far) it has been the most versatile and powerful.

Potential functions enable the application of optimization theory to the study of equilibria. More precisely, a potential function for a game is a real-valued function, defined on the set of possible outcomes of the game, such that the equilibria of the game are precisely the local optima of the potential function. This idea was first used to analyze selfish behavior in networks by Beckmann, McGuire, and Winsten [4], though similar ideas were used earlier in other contexts.

When a game admits a potential function, there are typically consequences for the existence, uniqueness, and inefficiency of equilibria. For example, suppose a game admits a potential function and either: (1) there are a finite number of distinct outcomes; or (2) the set of outcomes is compact and the potential function is continuous. In either case, the potential function achieves a global optimum, which is also a local optimum, and hence the game has at least one equilibrium. This is a much more elementary approach to establishing the existence of equilibria than traditional fixed-point proofs (e.g. [25]). Moreover, if the potential function has a unique local optimum, then the game has a unique equilibrium. Finally, if the potential function is "close to" the true objective function, then the equilibria that are global optima of the potential function have nearly-optimal objective function value, and are thus approximately efficient.

The power of the potential function approach might suggest that its applicability is limited. Fortunately, many important and natural classes of games admit well-behaved potential functions. To suggest what such functions look like, we briefly interpret some classical results about electric networks in terms of potential functions. Consider electrical current in a two-terminal network of resistors. By Kirchhoff's equations and Ohm's law, we can interpret this current as an "equilibrium", in the sense that it equalizes the voltage drop along all paths in the network between the two terminals. (View current as a large population of "selfish particles", each seeking out a path with minimum voltage drop.) On the other hand, Thomson's principle states that electrical current also minimizes the dissipated energy over all flow patterns that achieve the same total current. In other words, energy dissipation serves as a potential function for current in an electrical network. For further details and discussion, see Kelly [21] and Doyle and Snell [10].

**1.3. Survey overview.** Each of the next three sections introduces a model of selfish behavior in networks, and uses a potential function to bound the inefficiency of equilibria in the model. We focus on these three examples because they are simple, natural, and diverse enough to illustrate different aspects of potential function proof techniques. In order to emphasize the most important concepts and provide a number of self-contained proofs, we often discuss only special cases of more general models and results.

Section 2 discusses selfish routing networks, a model that generalizes the electrical networks of Subsection 1.2 and has been extensively studied by the transportation, networking, and theoretical computer science communities. Section 3 analyzes the performance of a well-studied distributed protocol for allocating resources to heterogeneous users. Section 4 bounds the inefficiency of equilibria in a model of selfish network design. Section 5 concludes.

## 2. Selfish routing and the price of anarchy

**2.1. The model.** In this section, we study the inefficiency of equilibria in the following model of noncooperative network routing. A *multicommodity flow network*, or simply a *network*, is a finite directed graph $G = (V, E)$, with vertex set $V$ and (directed) edge set $E$, together with a set $(s_1, t_1), \ldots, (s_k, t_k)$ of source-sink vertex pairs. We also call such pairs *commodities*. We denote the set of simple $s_i$-$t_i$ paths by $\mathcal{P}_i$, and always assume that this set is non-empty for each $i$. We allow the graph $G$ to contain parallel edges, and a vertex can participate in multiple source-sink pairs.

A *flow* in a network $G$ is a nonnegative vector indexed by the set $\mathcal{P} = \bigcup_{i=1}^{k} \mathcal{P}_i$. For a flow $f$ and a path $P \in \mathcal{P}_i$, we interpret $f_P$ as the amount of traffic of commodity $i$ that chooses the path $P$ to travel from $s_i$ to $t_i$. We use $r$ to denote a nonnegative vector of *traffic rates*, indexed by the commodities of $G$. A flow $f$ is *feasible* for $r$ if it routes all of the prescribed traffic: for each $i \in \{1, 2, \ldots, k\}$, $\sum_{P \in \mathcal{P}_i} f_P = r_i$.

We model the negative consequences of network congestion in the following simple way. For a flow $f$ in a network $G$ and an edge $e$ of $G$, let $f_e = \sum_{P \in \mathcal{P} : e \in P} f_P$ denote the total amount of traffic employing edge $e$. Each edge $e$ then has a nonnegative, continuous, and nondecreasing *cost function* $c_e$, which describes the cost incurred by traffic using the edge $e$ as a function of $f_e$. We call a triple of the form $(G, r, c)$ a *selfish routing network* or simply an *instance*.

Next we describe a notion of equilibrium in selfish routing networks – the expected outcome of "selfish routing". Define the cost of a path $P$ with respect to a flow $f$ as the sum of the costs of the constituent edges: $c_P(f) = \sum_{e \in P} c_e(f_e)$. Assuming that selfish traffic attempts to minimize its incurred cost, we obtain the following definition of a *Wardrop equilibrium* [38].

**Definition 2.1** ([38]). Let $f$ be a feasible flow for the instance $(G, r, c)$. The flow $f$ is a *Wardrop equilibrium* if, for every commodity $i \in \{1, 2, \ldots, k\}$ and every pair $P, \tilde{P} \in \mathcal{P}_i$ of $s_i$-$t_i$ paths with $f_P > 0$, $c_P(f) \leq c_{\tilde{P}}(f)$.

In Definition 2.1, we are implicitly assuming that every network user controls a negligible portion of the overall traffic, so that the actions of an individual user have essentially no effect on the network congestion. In the game theory literature, games with this property are called *nonatomic* [35]. Atomic variants of selfish routing have also been extensively studied (see e.g. [32]). We will study other types of atomic games in Sections 3 and 4.

**Example 2.2** (Pigou's example [27]). Consider the two-vertex, two-edge network shown in Figure 1. There is one commodity and the traffic rate is 1. Note that the lower edge is cheaper than the upper edge if and only if less than one unit of traffic uses it. There is thus a unique Wardrop equilibrium, with all traffic routed on the lower edge. In this flow, all traffic incurs one unit of cost.

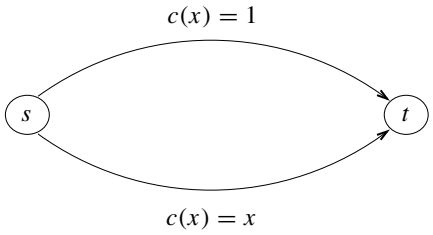



Figure 1. Pigou's example (Example 2.2).

Pigou's example already illustrates that equilibria can be inefficient. More specifically, note that routing half of the traffic on each of the two edges would produce a "better" flow: all of the traffic would incur at most one unit of cost, while half of the traffic would incur only $1/2$ units of cost.

The inefficiency of the Wardrop equilibrium in Example 2.2 arises from what is often called a *congestion externality* – a selfish network user accounts only for its own cost, and not for the costs that its decision imposes on others. The "better" routing of traffic in Example 2.2 is not a Wardrop equilibrium because a selfish network user routed on the upper edge would switch to the lower edge, indifferent to the fact that this switch (slightly) increases the cost incurred by a large portion of the population.

In Example 2.2, there is a unique Wardrop equilibrium. In Subsection 2.2 we will use a potential function to prove the following theorem, which states that Wardrop equilibria exist and are "essentially unique" in all selfish routing networks.

**Theorem 2.3** ([4]). *Let $(G, r, c)$ be an instance.*

(a) *The instance $(G, r, c)$ admits at least one Wardrop equilibrium.*

(b) *If $f$ and $\tilde{f}$ are Wardrop equilibria for $(G, r, c)$, then $c_e(f_e) = c_e(\tilde{f}_e)$ for every edge $e$.*

The Wardrop equilibrium in Example 2.2 is intuitively inefficient; we next quantify this inefficiency. We define our objective function, the *cost* of a flow, as the sum of the path costs incurred by traffic:

$$C(f) = \sum_{P \in \mathcal{P}} c_P(f) f_P = \sum_{e \in E} c_e(f_e) f_e. \tag{1}$$

The first equality in (1) is a definition; the second follows easily from the definitions. An *optimal flow* for an instance $(G, r, c)$ is feasible and minimizes the cost. Since cost functions are continuous and the set of feasible flows is compact, every instance admits an optimal flow. In Pigou's example (Example 2.2), the Wardrop equilibrium has cost 1, while routing half of the traffic on each edge yields an optimal flow with cost $3/4$.

**Definition 2.4** ([23], [26]). The *price of anarchy* $\rho(G, r, c)$ of an instance $(G, r, c)$ is

$$\rho(G, r, c) = \frac{C(f)}{C(f^*)},$$

where $f$ is a Wardrop equilibrium and $f^*$ is an optimal flow. The *price of anarchy* $\rho(\mathcal{I})$ of a non-empty set $\mathcal{I}$ of instances is $\sup_{(G,r,c) \in \mathcal{I}} \rho(G, r, c)$.

Definition 2.1 and Theorem 2.3(b) easily imply that all Wardrop equilibria have equal cost, and thus the price of anarchy of an instance is well defined unless there is a flow with zero cost. In this case, all Wardrop equilibria also have zero cost, and we define the price of anarchy of the instance to be 1.

**Example 2.5** (Nonlinear Pigou's example [34]). The inefficiency of the Wardrop equilibrium in Example 2.2 can be amplified with a seemingly minor modification to the network. Suppose we replace the previously linear cost function $c(x) = x$ on

the lower edge with the highly nonlinear one $c(x) = x^p$ for $p$ large (Figure 2). As in Example 2.2, the cost of the unique Wardrop equilibrium is 1. The optimal flow routes a small $\varepsilon$ fraction of the traffic on the upper edge and has cost $\varepsilon + (1-\varepsilon)^{p+1}$. Since this approaches 0 as $\varepsilon$ tends to 0 and $p$ tends to infinity, the price of anarchy of this selfish routing network grows without bound as $p$ grows large.

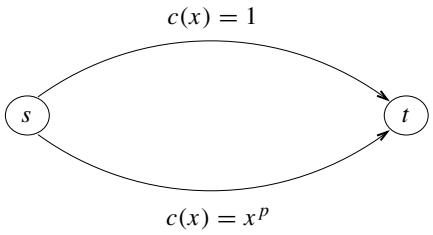



Figure 2. A nonlinear variant of Pigou's example (Example 2.5).

Example 2.5 demonstrates that the price of anarchy can be large in (very simple) networks with nonlinear cost functions. In Subsection 2.2 we use a potential function to show the converse: the price of anarchy is large *only* in networks with "highly nonlinear" cost functions.

**2.2. A potential function for wardrop equilibria.** We now show that Wardrop equilibria can be characterized as the minima of a potential function, and use this characterization to prove both Theorem 2.3 and upper bounds on the price of anarchy of selfish routing. To motivate this potential function, we first characterize the optimal flows of an instance.

Optimal flows for an instance $(G, r, c)$ minimize the cost (1) subject to linear flow feasibility constraints. Assume for the moment that for every edge $e$, the function $x \cdot c_e(x)$ is convex. The cost (1) is then a convex (separable) function, and we can apply the Karush–Kuhn–Tucker conditions (see e.g. [5]) to characterize its global minima. To state this characterization cleanly, assume further that all cost functions are differentiable, and let $c_e^*(x) = (x \cdot c_e(x))' = c_e(x) + x \cdot c_e'(x)$ denote the *marginal cost function* for the edge $e$. The KKT conditions then give the following.

**Proposition 2.6** ([4]). *Let $(G, r, c)$ be an instance such that, for every edge $e$, the function $x \cdot c_e(x)$ is convex and differentiable. Let $c_e^*$ denote the marginal cost function of the edge $e$. Then $f^*$ is an optimal flow for $(G, r, c)$ if and only if, for every commodity $i \in \{1, 2, \ldots, k\}$ and every pair $P, \tilde{P} \in \mathcal{P}_i$ of $s_i$-$t_i$ paths with $f_P > 0$, $c_P^*(f) \leq c_{\tilde{P}}^*(f)$.*

Comparing Definition 2.1 and Proposition 2.6, we discover that Wardrop equilibria and optimal flows are essentially the same thing, just with different sets of cost functions.

**Corollary 2.7.** *Let $(G, r, c)$ be an instance such that, for every edge $e$, the function $x \cdot c_e(x)$ is convex and differentiable. Let $c_e^*$ denote the marginal cost function of the edge $e$. Then $f^*$ is an optimal flow for $(G, r, c)$ if and only if it is a Wardrop equilibrium for $(G, r, c^*)$.*

To construct a potential function for Wardrop equilibria, we need to "invert" Corollary 2.7: of what function do Wardrop equilibria arise as global minima? The answer is simple: to recover Definition 2.1 as an optimality condition, we seek a function $h_e(x)$ for each edge $e$ – playing the previous role of $x \cdot c_e(x)$ – such that $h_e'(x) = c_e(x)$. Setting $h_e(x) = \int_0^x c_e(y)\,dy$ for each edge $e$ thus yields the desired potential function.

Precisely, call

$$\Phi(f) = \sum_{e \in E} \int_0^{f_e} c_e(x)\,dx \tag{2}$$

the *potential function* for an instance $(G, r, c)$. Analogously to Corollary 2.7, the following proposition holds.

**Proposition 2.8** ([4]). *Let $(G, r, c)$ be an instance. A flow feasible for $(G, r, c)$ is a Wardrop equilibrium if and only if it is a global minimum of the corresponding potential function $\Phi$ given in* (2)*.*

**Remark 2.9.** Thomson's principle for electrical networks (Subsection 1.2) can be viewed as the special case of Proposition 2.8 for single-commodity flow networks with linear cost functions (of the form $c_e(x) = a_e x$).

Theorem 2.3 now follows easily.

*Proof of Theorem* 2.3 (Sketch). Since cost functions are continuous and the set of feasible flows is compact, part (a) of the theorem follows immediately from Proposition 2.8 and Weierstrass's Theorem. Since cost functions are nondecreasing, the potential function $\Phi$ in (2) is convex; moreover, the set of feasible flows is convex. Part (b) of the theorem now follows from routine convexity arguments. □

Much more recently, the potential function (2) has been used to upper bound the price of anarchy of selfish routing. The intuition behind this connection is simple: if Wardrop equilibria exactly optimize a potential function (2) that is a good approximation of the objective function (1), then they should also be approximately optimal. Formally, we have the following.

**Theorem 2.10** ([34]). *Let $(G, r, c)$ be an instance, and suppose that $x \cdot c_e(x) \leq \gamma \cdot \int_0^x c_e(y)\,dy$ for all $e \in E$ and $x \geq 0$. Then the price of anarchy $\rho(G, r, c)$ is at most $\gamma$.*

*Proof.* Let $f$ and $f^*$ be a Wardrop equilibrium and an optimal flow for $(G, r, c)$, respectively. Since cost functions are nondecreasing, the cost of a flow (1) is always

at least its potential function value (2). The hypothesis ensures that the cost of a flow is at most $\gamma$ times its potential function value. The theorem follows by writing

$$C(f) \leq \gamma \cdot \Phi(f) \leq \gamma \cdot \Phi(f^*) \leq \gamma \cdot C(f^*),$$

with the second inequality following from Proposition 2.8. □

Theorem 2.10 implies that the price of anarchy of selfish routing is large only in networks with "highly nonlinear" cost functions. For example, if $c_e$ is a polynomial function with degree at most $p$ and nonnegative coefficients, then $x \cdot c_e(x) \leq (p+1)\int_0^x c_e(y)\,dy$ for all $x \geq 0$. Applying Theorem 2.10, we find that the price of anarchy in networks with cost functions that are polynomials with nonnegative coefficients grows at most linearly with the degree bound $p$.

**Corollary 2.11** ([34]). *If $(G, r, c)$ is an instance with cost functions that polynomials with nonnegative coefficients and degree at most $p$, then $\rho(G, r, c) \leq p + 1$.*

This upper bound is nearly matched by Example 2.5. (The upper and lower bounds differ by roughly a $\ln p$ multiplicative factor.) Qualitatively, Example 2.5 and Corollary 2.11 imply that a large price of anarchy can be caused by highly nonlinear cost functions, but not by complex network topologies or by a large number of commodities.

**2.3. An optimal bound on the price of anarchy.** We have established that the price of anarchy of selfish routing depends on the "degree of nonlinearity" of the network cost functions. However, even in the simple case of polynomial cost functions, there is gap between the lower bound on the price of anarchy provided by Example 2.5 and the upper bound of Theorem 2.10. We conclude this section by showing how a different analysis, which can be regarded as a more "global" application of potential function ideas, provides a tight bound on the price of anarchy for essentially every set of allowable cost functions.

We first formalize a natural lower bound on the price of anarchy based on "Pigou-like examples".

**Definition 2.12** ([7], [31]). Let $\mathcal{C}$ be a nonempty set of cost functions. The *Pigou bound* $\alpha(\mathcal{C})$ for $\mathcal{C}$ is

$$\alpha(\mathcal{C}) = \sup_{c \in \mathcal{C}} \sup_{x, r \geq 0} \frac{r \cdot c(r)}{x \cdot c(x) + (r - x)c(r)}, \tag{3}$$

with the understanding that $0/0 = 1$.

The point of the Pigou bound is that it lower bounds the price of anarchy in instances with cost functions in $\mathcal{C}$.

**Proposition 2.13.** *Let $\mathcal{C}$ be a set of cost functions that contains all of the constant cost functions. Then $\rho(\mathcal{C}) \geq \alpha(\mathcal{C})$.*

*Proof.* Fix a choice of $c \in \mathcal{C}$ and $x, r \geq 0$. We can complete the proof by exhibiting a selfish routing network with cost functions in $\mathcal{C}$ and price of anarchy at least $c(r)r/[c(x)x + (r-x)c(r)]$. Since $c$ is nondecreasing, this expression is at least 1 if $x \geq r$; we can therefore assume that $x < r$.

Let $G$ be a two-vertex, two-edge network as in Figure 1. Give the lower edge the cost function $c_1(y) = c(y)$ and the upper edge the constant cost function $c_2(y) = c(r)$. By assumption, both of these cost functions lie in $\mathcal{C}$. Set the traffic rate to $r$. Routing all of the traffic on the lower edge yields a Wardrop equilibrium with cost $c(r)r$. Routing $x$ units of traffic on the lower edge and $r - x$ units of traffic on the upper edge gives a feasible flow with cost $[c(x)x + (r-x)c(r)]$. The price of anarchy in this instance is thus at least $c(r)r/[c(x)x + (r-x)c(r)]$, as desired □

Proposition 2.13 holds more generally for every set $\mathcal{C}$ of cost functions that is *inhomogeneous* in the sense that $c(0) > 0$ for some $c \in \mathcal{C}$ [31].

We next show that, even though the Pigou bound is based only on Pigou-like examples, it is also an upper bound on the price of anarchy in general multicommodity flow networks. The proof requires the following *variational inequality* characterization of Wardrop equilibria, first noted by Smith [36].

**Proposition 2.14** ([36]). *A flow $f$ feasible for $(G, r, c)$ is a Wardrop equilibrium if and only if*

$$\sum_{e \in E} c_e(f_e) f_e \leq \sum_{e \in E} c_e(f_e) f_e^*$$

*for every flow $f^*$ feasible for $(G, r, c)$.*

Proposition 2.14 can be derived as an optimality condition for minimizers of the potential function (2), or can be proved directly using Definition 2.1.

We now show that the Pigou bound is tight.

**Theorem 2.15** ([7], [31]). *Let $\mathcal{C}$ be a set of cost functions and $\alpha(\mathcal{C})$ the Pigou bound for $\mathcal{C}$. If $(G, r, c)$ is an instance with cost functions in $\mathcal{C}$, then*

$$\rho(G, r, c) \leq \alpha(\mathcal{C}).$$

*Proof.* Let $f^*$ and $f$ be an optimal flow and a Wardrop equilibrium, respectively, for an instance $(G, r, c)$ with cost functions in the set $\mathcal{C}$. The theorem follows by writing

$$\begin{aligned} C(f^*) &= \sum_{e \in E} c_e(f_e^*) f_e^* \\ &\geq \frac{1}{\alpha(\mathcal{C})} \sum_{e \in E} c_e(f_e) f_e + \sum_{e \in E} (f_e^* - f_e) c_e(f_e) \\ &\geq \frac{C(f)}{\alpha(\mathcal{C})}, \end{aligned}$$

where the first inequality follows from Definition 2.12 and the second from Proposition 2.14. □

Different, more recent proofs of Theorem 2.15 can be found in [8], [37].

Proposition 2.13 and Theorem 2.15 establish the qualitative statement that, for essentially every fixed restriction on the allowable network cost functions, the price of anarchy is maximized by Pigou-like examples. Determining the largest-possible price of anarchy in Pigou-like examples (i.e., the Pigou bound) is a tractable problem in many cases. For example, it is precisely $4/3$ when $\mathcal{C}$ is the affine functions [34], and more generally is $[1 - p \cdot (p+1)^{-(p+1)/p}]^{-1} \approx p/\ln p$ when $\mathcal{C}$ is the set of polynomials with degree at most $p$ and nonnegative coefficients [31]. In these cases, the maximum price of anarchy (among all multicommodity flow networks) is achieved by the instances in Examples 2.2 and 2.5. For further examples, see [7], [31].

For much more on topics related to the price of anarchy of selfish routing, including many extensions and generalizations of the results described in this section, see [32], [33] and the references therein.

## 3. Efficiency loss in resource allocation protocols

We next study the performance of a protocol for allocating resources to heterogeneous users. While there are a number of conceptual differences between this model and the selfish routing networks of Section 2, the inefficiency of equilibria in these models can be analyzed in a similar way.

**3.1. The model.** We consider a single divisible resource – the capacity of a single network link, say – to be allocated to a finite number $n > 1$ of competing users. These users are *heterogeneous* in the sense that different users can have different values for capacity. We model this by giving each user $i$ a nonnegative real-valued *utility function* $U_i$ that expresses this user's value for a given amount of capacity. We assume that each $U_i$ is concave, strictly increasing, and continuously differentiable. A *resource allocation game* is defined by the $n$ utility functions $U_1, \ldots, U_n$ and the link capacity $C > 0$.

An *allocation* for a resource allocation game is a nonnegative vector $(x_1, \ldots, x_n)$ with $\sum_{i=1}^n x_i = C$. We study the following protocol for allocating capacity. Each user $i$ submits a nonnegative *bid* $b_i$ for capacity. The protocol allocates capacity in proportion to bids, with

$$x_i = \frac{b_i}{\sum_{j=1}^n b_j} \cdot C \tag{4}$$

units of capacity allocated to user $i$. The *payoff* $Q_i$ to a user $i$ is defined as its value for the capacity it receives, minus the bid that it made (and presumably now has to pay):

$$Q_i(b_1, \ldots, b_n) = U_i(x_i) - b_i = U_i\left(\frac{b_i}{\sum_{j=1}^n b_j} \cdot C\right) - b_i.$$

Assume that if all users bid zero, then all users receive zero payoff.

An *equilibrium* is then a bid vector in which each user bids optimally, given the bids of the other users. To state this precisely, we use the notation $b_{-i} = (b_1, b_2, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n)$ to denote the bids of the users other than $i$, and sometimes write $(b_i, b_{-i})$ for a bid vector $(b_1, \ldots, b_n)$.

**Definition 3.1.** A bid vector $(b_1, \ldots, b_n)$ is an *equilibrium* of the resource allocation game $(U_1, \ldots, U_n, C)$ if for every user $i \in \{1, 2, \ldots, n\}$,

$$Q_i(b_i, b_{-i}) = \sup_{\tilde{b}_i \geq 0} Q_i(\tilde{b}_i, b_{-i}). \tag{5}$$

One easily checks that in every equilibrium, at least two users submit strictly positive bids.

While equilibria are most naturally defined for bid vectors, we will be interested in the quality of the corresponding allocations. An *equilibrium allocation* is an allocation $(x_1, \ldots, x_n)$ induced by an equilibrium bid vector – i.e., there is an equilibrium $(b_1, \ldots, b_n)$ such that (4) holds for each user $i$. We next give a characterization of equilibrium allocations that will be crucial for designing a potential function for resource allocation games.

First, a simple calculation shows that concavity of the utility function $U_i$ (in $x_i$) implies strict concavity of the payoff function $Q_i$ (in $b_i$) for every fixed vector $b_{-i}$ with at least one strictly positive component. Similarly, the latter function is continuously differentiable for each such fixed $b_{-i}$. We can therefore characterize solutions to (5) via standard first-order optimality conditions, which yields the following.

**Proposition 3.2** ([16], [20]). *Let $(U_1, \ldots, U_n, C)$ be a resource allocation game and $(b_1, \ldots, b_n)$ a bid vector with at least two strictly positive bids. Let $B = \sum_{j=1}^{n} b_j$ denote the sum of the bids. This bid vector is an equilibrium if and only if*

$$U_i'\left(\frac{b_i}{B} \cdot C\right)\left(1 - \frac{b_i}{B}\right) \leq \frac{B}{C}$$

*for every user $i \in \{1, 2, \ldots, n\}$, with equality holding whenever $b_i > 0$.*

Reformulating Proposition 3.2 in terms of allocations gives the following corollary (cf., Definition 2.1).

**Corollary 3.3** ([16], [20]). *Let $(U_1, \ldots, U_n, C)$ be a resource allocation game. An allocation $(x_1, \ldots, x_n)$ is an equilibrium if and only if for every pair $i$, $j \in \{1, 2, \ldots, n\}$ of users with $x_i > 0$,*

$$U_i'(x_i)\left(1 - \frac{x_i}{C}\right) \geq U_j'(x_j)\left(1 - \frac{x_j}{C}\right).$$

*Proof.* The "only if" direction follows easily from Proposition 3.2 and equation (4). For the "if" direction, suppose $(x_1, \ldots, x_n)$ satisfies the stated condition. There is then a scalar $\lambda \geq 0$ such that $U_i'(x_i)[1 - (x_i/C)] \leq \lambda$ for all users $i$, with equality holding whenever $x_i > 0$. Setting $b_i = \lambda x_i$ for each $i$ yields a bid vector that meets the equilibrium condition in Proposition 3.2. □

**Example 3.4** ([20]). Consider a resource allocation game in which the capacity $C$ is 1, one user has the utility function $U_1(x_1) = 2x_1$, and the other $n-1$ users have the utility function $U_i(x_i) = x_i$. Corollary 3.3 implies that in the unique equilibrium allocation, the first user receives $\frac{1}{2} + \varepsilon$ units of capacity, while each of the other $n-1$ users receive $\delta$ units of capacity (with $\varepsilon, \delta \to 0$ as $n \to \infty$). In this allocation, $U_i'(x_i)(1-x_i)$ is the same for each user $i$, and is slightly less than 1. The corresponding equilibrium bid vector is roughly the same as the equilibrium allocation vector.

In the next subsection, we use a potential function to show that every resource allocation game has a unique equilibrium allocation.

We claim that the equilibrium allocation in Example 3.4 is suboptimal. As in the previous section, we formalize this claim by introducing an objective function and studying the price of anarchy. We define the *efficiency* of an allocation $(x_1, \ldots, x_n)$ of a resource allocation game to be the sum of the users' utilities:

$$\mathcal{E}(x_1, \ldots, x_n) = \sum_{i=1}^{n} U_i(x_i). \tag{6}$$

An *optimal* allocation has the maximum-possible efficiency.

The *price of anarchy* of a resource allocation game is the ratio $\mathcal{E}(x)/\mathcal{E}(x^*)$, where $x$ is the equilibrium allocation and $x^*$ is an optimal allocation. Note that the price of anarchy of such a game is at most 1. In Example 3.4, the optimal allocation gives all of the capacity to the first user and has efficiency 2. The equilibrium allocation has efficiency approaching $3/2$ as $n \to \infty$; the price of anarchy can therefore be arbitrarily close to $3/4$ in this family of examples.

Why does inefficiency arise in Example 3.4? First, note that if the first user is the only one bidding a strictly positive amount (leading to the optimal allocation), then the bid vector cannot be an equilibrium: the first user can bid a smaller positive amount and continue to receive all of the capacity. A similar argument holds whenever the first user's bid comprises a sufficiently large fraction of the sum of the users' bids: if the first user lowers its bid, its allocation diminishes, but the price it pays per unit of bandwidth decreases by a large enough amount to increase its overall payoff. This intuition is mathematically reflected in Corollary 3.3 in the term $U_i'(x_i)(1 - x_i)$ – the marginal benefit of increased capacity to a user becomes increasing tempered as its allocation grows. Inefficiency thus arises in Example 3.4 because of "market power" – the fact that the actions of a single user have significant influence over the effective price of capacity. Indeed, resource allocation games were initially studied by Kelly [22] under the assumption that no users have nontrivial market power. Under this assumption, equilibria are fully efficient – i.e., the price of anarchy is always 1 [22]. See [19, §1.3–1.4] for further discussion.

**Remark 3.5.** Selfish routing networks and resource allocation games differ in a number of ways. In the former, there is a continuum of selfish network users that each have a finite set of strategies (paths); in the latter, there is a finite set of users, each with a

continuum of strategies (bids). In selfish routing, the objective is cost minimization; in resource allocation, it is efficiency maximization. Finally, and perhaps most fundamentally, inefficiency appears to arise for different reasons in the two models. Recall that in selfish routing networks, inefficiency stems from congestion externalities (see the discussion following Example 2.2). Example 3.4 shows that market power is the culprit behind inefficient equilibria in resource allocation games. Despite all of these conceptual differences, the next two subsections show that the inefficiency of equilibria can be quantified in the two models via remarkably similar analyses.

**3.2. A potential function for equilibria.** As in Subsection 2.2, our first step toward constructing a potential function for equilibrium allocations is to characterize optimal allocations. Since efficiency (6) is a separable concave function, a straightforward application of first-order optimality conditions yields the following.

**Proposition 3.6.** *Let $(U_1, \ldots, U_n, C)$ be a resource allocation game. An allocation $(x_1, \ldots, x_n)$ is optimal if and only if for every pair $i, j \in \{1, 2, \ldots, n\}$ of users with $x_i > 0$, $U_i'(x_i) \geq U_j'(x_j)$.*

Given the near-identical characterizations of equilibrium and optimal allocations in Corollary 3.3 and Proposition 3.6, respectively, we again ask: of what function does an equilibrium allocation arise as the global maximum? To recover Corollary 3.3 as an optimality condition, we seek a function $H_i$ for each user $i$ such that $H_i'(x_i) = U_i'(x_i)[1 - (x_i/C)]$ for all $x_i \geq 0$. Setting $H_i(x_i) = U_i(x_i)[1 - (x_i/C)] + \left[\int_0^{x_i} U_i(y)\,dy\right]/C$ thus yields the desired potential function. Precisely, for the resource allocation game $(U_1, \ldots, U_n, C)$, define

$$\Phi_{RA}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \hat{U}_i(x_i), \tag{7}$$

where

$$\hat{U}_i(x_i) = \left(1 - \frac{x_i}{C}\right) \cdot U_i(x_i) + \frac{x_i}{C} \cdot \left(\frac{1}{x_i} \int_0^{x_i} U_i(y)\,dy\right). \tag{8}$$

A simple calculation shows that each function $\hat{U}_i$ is strictly concave, increasing, and continuously differentiable. Regarding $(\hat{U}_1, \ldots, \hat{U}_n, C)$ as a resource allocation game, applying Proposition 2.6 to it, and appealing to Corollary 3.3 formalizes the fact that $\Phi_{RA}$ is a potential function.

**Proposition 3.7** ([16], [20]). *An allocation of the game $(U_1, \ldots, U_n, C)$ is an equilibrium allocation if and only if it is a global maximum of the corresponding potential function $\Phi_{RA}$.*

Existence and uniqueness of equilibrium allocations follow immediately.

**Proposition 3.8** ([16], [20]). *In every resource allocation game, there is a unique equilibrium allocation.*

*Proof.* Existence follows from Proposition 3.7 and the facts that the potential function (7) is continuous and the set of all allocations is compact. Uniqueness follows from Proposition 3.7 and the fact that the potential function (7) is strictly concave. □

Proposition 3.7 also has consequences for the price of anarchy in resource allocation games. To see why, note that the value of $\hat{U}_i(x_i)$ in (8) can be viewed as a weighted average of two quantities – the "true utility" $U_i(x_i)$ and the "average utility" $\left[\int_0^{x_i} U_i(y)\,dy\right]/x_i$. Since $U_i$ is increasing, the latter quantity can only underestimate the utility $U_i(x_i)$, and hence $\hat{U}_i(x_i) \leq U_i(x_i)$ for all $i$ and $x_i \geq 0$. On the other hand, since $U_i$ is nonnegative and concave, the average utility between 0 and $x_i$ is at least half of the utility $U_i(x_i)$ at $x_i$. Thus $\hat{U}_i(x_i) \geq U_i(x_i)/2$ for all $i$ and $x_i \geq 0$. It follows that

$$\mathcal{E}(x_1, \ldots, x_n) \geq \Phi_{RA}(x_1, \ldots, x_n) \geq \mathcal{E}(x_1, \ldots, x_n)/2$$

for every allocation $(x_1, \ldots, x_n)$. Following the proof of Theorem 2.10 now gives a lower bound of $1/2$ on the price of anarchy in resource allocation games.

**Theorem 3.9** ([20]). *In every resource allocation game, the price of anarchy is at least* $1/2$.

**3.3. An optimal bound on the price of anarchy.** There is a gap between the lower bound of $1/2$ on the price of anarchy given in Theorem 3.9 and the upper bound of $3/4$ that is achieved (in the limit) in Example 3.4. As in Subsection 3.3, an optimal (lower) bound can be obtained by leveraging the potential function characterization of equilibria (Proposition 3.7) in a less crude way. Our argument will again be based on a "variational inequality", which can be derived directly from Corollary 3.3 or viewed as a first-order optimality condition for the potential function (7).

**Proposition 3.10.** *Let $(U_1, \ldots, U_n, C)$ be a resource allocation game. For each user $i$, define the modified utility function $\hat{U}_i$ as in* (8)*. An allocation $\hat{x}$ is an equilibrium for $(U_1, \ldots, U_n, C)$ if and only if*

$$\sum_{i=1}^{n} \hat{U}_i'(\hat{x}_i)\hat{x}_i \geq \sum_{i=1}^{n} \hat{U}_i'(\hat{x}_i)x_i$$

*for every feasible allocation $x$.*

Next is the analogue of the Pigou bound (Definition 2.12) for resource allocation games. This definition is primarily motivated by the upper bound on the price of anarchy provided by Example 3.4; we state it in a form that also permits easy application of Proposition 3.10 in the proof of Lemma 3.13 below.

**Definition 3.11.** Let $\mathcal{U}$ denote the set of real-valued, nonnegative, strictly increasing, continuously differentiable, and concave (utility) functions. Define the *JT bound* $\beta$ by

$$\beta = \inf_{U \in \mathcal{U}} \inf_{C > 0} \inf_{0 \leq \hat{x}, x^* \leq C} \frac{U(\hat{x}) + \hat{U}'(\hat{x})(x^* - \hat{x})}{U(x^*)}, \tag{9}$$

where $\hat{U}$ is defined as in (8), as a function of $U$ and $C$.

In the rest of this section, we show that the JT bound is exactly the worst price of anarchy occurring in resource allocation games, and explicitly compute the bound.

**Lemma 3.12.** *For every $\varepsilon > 0$, there is a resource allocation game with price of anarchy at most $\beta + \varepsilon$, where $\beta$ is the JT bound.*

**Lemma 3.13.** *In every resource allocation game, the price of anarchy is at least the JT bound $\beta$.*

**Lemma 3.14.** *The JT bound $\beta$ is exactly $3/4$.*

Lemmas 3.12–3.14 give an explicit optimal bound on the price of anarchy in resource allocation games.

**Theorem 3.15** ([20]). *In every resource allocation game, the price of anarchy is at least $3/4$. Moreover, this bound is tight.*

We now prove Lemmas 3.12–3.14 in turn.

*Proof of Lemma* 3.12. Fix a choice of a utility function $U$, a capacity $C > 0$, and values for $\hat{x}, x^* \in [0, C]$. We aim to exhibit a resource allocation game with price of anarchy (arbitrarily close to)

$$\frac{U(\hat{x}) + \hat{U}'(\hat{x})(x^* - \hat{x})}{U(x^*)}. \tag{10}$$

Recall from (8) that $\hat{U}'(\hat{x}) = U'(\hat{x}) \cdot [1 - (\hat{x}/C)]$. A calculation shows that (10) is at least 1 if $\hat{x} \geq x^*$, so we can assume that $\hat{x} < x^*$. Since (10) is decreasing in $C$, we can assume that $C = x^*$.

Define a resource allocation game in which the capacity is $C$, the first user has the utility function $U_1(x_1) = U(x_1)$, and the other $n - 1$ users each have the linear utility function $U_i(x_i) = \hat{U}'(\hat{x}) \cdot x_i$. Giving all of the capacity to the first user is a feasible allocation with efficiency $U_1(C) = U(x^*)$. Arguing as in Example 3.4, the equilibrium allocation has efficiency approaching $U_1(\hat{x}) + (C - \hat{x}) \cdot \hat{U}'(\hat{x}) = U(\hat{x}) + \hat{U}'(\hat{x})(x^* - \hat{x})$ as the number $n$ of users tends to infinity. The price of anarchy in this family of instances thus tends to (at most) the expression in (10) as $n \to \infty$, completing the proof. □

*Proof of Lemma* 3.13. Let $(U_1, \ldots, U_n, C)$ be a resource allocation game. Let $x^*$ and $\hat{x}$ denote optimal and equilibrium allocations, respectively. Define the modified utility function $\hat{U}_i$ as in (8). The lemma follows by writing

$$\begin{aligned}\sum_{i=1}^{n} U_i(x_i^*) &\leq \sum_{i=1}^{n} \left[\frac{1}{\beta}\big(U_i(\hat{x}_i) + \hat{U}_i'(\hat{x}_i)(x_i^* - \hat{x}_i)\big)\right] \\ &\leq \frac{1}{\beta} \sum_{i=1}^{n} U_i(\hat{x}_i),\end{aligned}$$

where the first inequality follows from Definition 3.11 and the second from Proposition 3.10. □

*Proof of Lemma* 3.14. Setting $U$ to the identity function, $\hat{x} = 1/2$, and $C = x^* = 1$ shows that the JT bound is at most $3/4$. Now fix arbitrary choices of $U$, $C$, and $\hat{x}, x^* \in [0, C]$. We need to show that (10) is at least $3/4$. As in the proof of Lemma 3.12, we can assume that $\hat{x} < x^* = C$. We can then write

$$\begin{aligned}
U(\hat{x}) + \hat{U}'(\hat{x})(x^* - \hat{x}) &= U(\hat{x}) + \left(1 - \frac{\hat{x}}{x^*}\right)U'(\hat{x})(x^* - \hat{x}) \\
&\geq U(\hat{x}) + \left(1 - \frac{\hat{x}}{x^*}\right)\big(U(x^*) - U(\hat{x})\big) \\
&= \left(\frac{\hat{x}}{x^*}\right) \cdot U(\hat{x}) + \left(1 - \frac{\hat{x}}{x^*}\right) \cdot U(x^*) \\
&\geq \left(\frac{\hat{x}}{x^*}\right)^2 \cdot U(x^*) + \left(1 - \frac{\hat{x}}{x^*}\right) \cdot U(x^*) \\
&\geq \frac{3}{4} \cdot U(x^*),
\end{aligned}$$

where the first equality follows from the definition of $\hat{U}$ in (8), the first and second inequalities follow from the concavity and nonnegativity of $U$, and the final inequality follows from the fact that the function $y^2 - y + 1$ is uniquely minimized when $y = 1/2$. The proof is complete. □

**Remark 3.16.** The original proof of Theorem 3.15 is fairly different than the one given here. Specifically, Johari and Tsitsiklis [20] first show that the price of anarchy is minimized in games in which all users have linear utility functions, and then explicitly determine a worst-case example (the same as Example 3.4) by analyzing a linear program. We instead presented the proof above to further highlight the connections between resource allocation games and selfish routing networks.

Despite the numerous common features in our analyses of the price of anarchy in selfish routing networks and in resource allocation games, the precise relationship between the two models is not completely understood. In particular, we lack a unifying analysis of the price of anarchy in the two models.

**Open Question 1.** Find a compelling generalization of selfish routing networks and resource allocation games in which the price of anarchy can be analyzed in a uniform way. Ideally, such a generalization would unify Theorems 2.15 and 3.15, and would also apply to several of the more general classes of games described in [19], [32].

As with selfish routing networks, we have only scratched the surface of the literature on the price of anarchy in resource allocation games. For much more on the subject, including generalizations of these games to general networks, see Johari and Tsitsiklis [20] and Johari [19].

## 4. The price of stability in network design games

Our final class of games is a model of network design with selfish users. These games share some features with selfish routing networks, but also differ in a few fundamental respects.

**4.1. The model.** In this section we study *Shapley network design games*, first proposed by Anshelevich et al. [1]. The game occurs in a directed graph $G = (V, E)$, in which each edge $e \in E$ has a fixed nonnegative cost $c_e$. There is a finite set of $k$ selfish players, and each player $i \in \{1, 2, \ldots, k\}$ is identified with a source-sink vertex pair $(s_i, t_i)$. Let $\mathcal{P}_i$ denote the set of simple $s_i$-$t_i$ paths.

Each player $i$ chooses a path $P_i \in \mathcal{P}_i$ from its source to its destination. This creates a network $\left(V, \bigcup_i P_i\right)$, and we define the *cost* of this outcome as

$$c(P_1, \ldots, P_k) = \sum_{e \in \bigcup_i P_i} c_e. \tag{11}$$

We assume that this cost is shared among the players in the following way. First, if edge $e$ lies in $f_e$ of the chosen paths, then each player choosing such a path pays a proportional share $\pi_e = c_e/f_e$ of the cost. The overall cost $c_i(P_1, \ldots, P_k)$ to player $i$ is then the sum $\sum_{e \in P_i} \pi_e$ of these proportional shares. Selfish players naturally attempt to minimize their incurred cost.

We next define our notion of equilibria for Shapley network design games. In contrast to selfish routing networks and resource allocation games, these network design games are finite games – there is a finite set of players, each with a finite set of strategies. This is the classical setting for *Nash equilibria* [25]. As in Definition 3.1, we use $P_{-i}$ to denote the vector of strategies chosen by the players other than $i$.

**Definition 4.1.** An outcome $(P_1, \ldots, P_k)$ of a Shapley network design game is a (*pure-strategy*) *Nash equilibrium* if for every player $i$,

$$c_i(P_i, P_{-i}) = \min_{\tilde{P}_i \in \mathcal{P}_i} c_i(\tilde{P}_i, P_{-i}).$$

In a pure-strategy Nash equilibrium, every player chooses a single strategy. In a *mixed-strategy* Nash equilibrium, a player can randomize over several strategies. We will not discuss mixed-strategy Nash equilibria in this survey, though the price of anarchy of such equilibria has been studied in different models (see e.g. [3], [23]).

**Example 4.2** ([2]). Consider the network shown in Figure 3. There are $k$ players, each with the same source $s$ and sink $t$. The edge costs are $k$ and $1 + \varepsilon$, where $\varepsilon > 0$ is arbitrarily small. In the minimum-cost outcome, all players choose the lower edge. This outcome is also a Nash equilibrium. On the other hand, suppose all of the players choose the upper edge. Each player $i$ then incurs cost 1, and if player $i$ deviates to the lower edge it pays the full cost of $1 + \varepsilon$. This outcome is thus a second Nash equilibrium, and it has cost $k$.

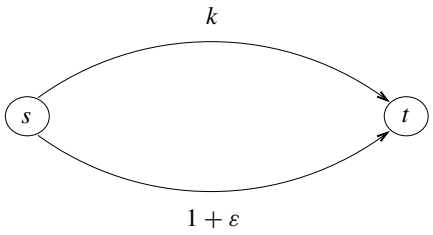



Figure 3. Multiple Nash equilibria in Shapley network design games (Example 4.2).

Example 4.2 shows that Shapley network design games are more ill-behaved than selfish routing networks and resource allocation games in a fundamental respect: there can be multiple equilibria, and different equilibria can have very different objective function values. (Cf., Theorem 2.3 and Proposition 3.8.) The definition of the price of anarchy is ambiguous in games with multiple equilibria – we would like to quantify the inefficiency of an equilibrium, but of which one?

The price of anarchy is historically defined as the ratio between the objective function value of the *worst* equilibrium and that of an optimal solution [23], [26]. This definition is natural from the perspective of worst-case analysis. In Example 4.2, the price of anarchy is (arbitrarily close to) $k$. It is also easy to show that the price of anarchy in every Shapley network design game is at most $k$.

In this section, we instead focus on the ratio between the cost of the *best* Nash equilibrium of a Shapley network design game and that of an optimal solution. This measure is called the *price of stability* [1]. Our motivation is twofold. First, as Example 4.2 shows, the price of anarchy is large and trivial to determine. Second, the price of stability has a reasonably natural interpretation in network design games – if we envision the network as being designed by a central authority for subsequent use by selfish players, then the best Nash equilibrium is an obvious solution to propose. In this sense, the price of stability measures the necessary degradation in solution quality caused by imposing the game-theoretic constraint of stability. See [1], [2], [6], [7] for further discussion and examples of the price of stability.

The price of stability in Example 4.2 is 1. We conclude this subsection with an example showing that this is not always the case.

**Example 4.3** ([1]). Consider the network shown in Figure 4. There are $k$ players, all with the same sink $t$, and $\varepsilon > 0$ is arbitrarily small. For each $i \in \{1, 2, \ldots, k\}$, the edge $(s_i, t)$ has cost $1/i$. In the minimum-cost outcome, each player $i$ chooses the path $s_i \to v \to t$ and the cost is $1 + \varepsilon$. This is not a Nash equilibrium, as player $k$ can decrease its cost from $(1 + \varepsilon)/k$ to $1/k$ by switching to the direct path $s_k \to t$. More generally, this direct path is a *dominant strategy* for the $k$th player – it is the minimum-cost strategy, independent of the paths chosen by the other players. It follows that in every Nash equilibrium, the $k$th player selects its direct path. Arguing inductively

about the players $k-1, k-2, \ldots, 1$, we find that the unique Nash equilibrium is the outcome in which each player $i$ chooses its direct path $s_i \to t$ to the sink. The cost of this outcome is exactly the $k$th harmonic number $\mathcal{H}_k = \sum_{i=1}^{k}(1/i)$, which is roughly $\ln k$. The price of stability can therefore be (arbitrarily close to) $\mathcal{H}_k$ in Shapley network design games.

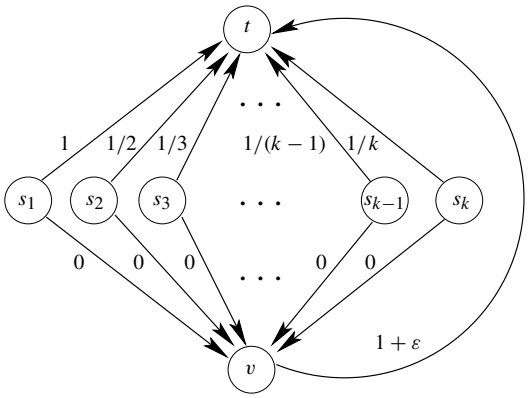



Figure 4. The price of stability in Shapley network design games can be at least $\mathcal{H}_k$ (Example 4.3).

**4.2. A potential function for Nash equilibria.** In this subsection we use a potential function to prove the existence of pure-strategy Nash equilibria and upper bound the price of stability in Shapley network design games. Recall that for both selfish routing networks and resource allocation games, we designed potential functions using a characterization of optimal solutions as a guide (see Propositions 2.6 and 3.6). In Shapley network design games, computing an optimal solution is an NP-hard network design problem [15], and we cannot expect to find an analogous characterization.

There are two ways that Shapley network design games differ from selfish routing networks that prevent the characterization of optimal solutions (Proposition 2.6) from carrying over. First, there are a finite number of players in the former model, and a continuum of players in the latter model. Second, cost functions in selfish routing networks are nondecreasing, whereas Shapley network design games effectively have cost functions that are *decreasing* in the "congestion" – if $x \geq 1$ players use an edge $e$ with fixed cost $c_e$, then the per-player cost on that edge is $c_e/x$.

On the bright side, the potential function (2) for selfish routing networks is easily modified to account for these two differences. First, note that this function $\Phi$ remains well-defined for decreasing cost functions. Second, passing from an infinite player set to a finite one merely involves changing the integrals in (2) to sums. This motivates

the following proposal for a potential function for a Shapley network design game:

$$\Phi_{ND}(P_1, \ldots, P_k) = \sum_{e \in E} \sum_{i=1}^{f_e} \frac{c_e}{i}, \tag{12}$$

where $f_e$ denotes the number of paths $P_i$ that include edge $e$. While equilibria in selfish routing networks and resource allocation games can be characterized as the *global* optima of their respective potential functions (2) and (7), we will see that the Nash equilibria of a Shapley network design game are characterized as the *local* optima of the potential function (12). This idea is originally due to Rosenthal [29], [30], who also considered the broader context of "atomic congestion games".

The next lemma, which is crucial for the rest of this section, states that the potential function "tracks" the change in cost experienced by a deviating player.

**Lemma 4.4** ([1], [30]). *Let $(G, c)$ denote a Shapley network design game with $k$ players and $\Phi_{ND}$ the corresponding potential function* (12)*. Let $i \in \{1, 2, \ldots, k\}$ be a player, and let $(P_i, P_{-i})$ and $(\tilde{P}_i, P_{-i})$ denote two outcomes that differ only in the strategy chosen by the $i$th player. Then*

$$c_i(\tilde{P}_i, P_{-i}) - c_i(P_i, P_{-i}) = \Phi_{ND}(\tilde{P}_i, P_{-i}) - \Phi_{ND}(P_i, P_{-i}). \tag{13}$$

*Proof.* Let $f_e$ denote the number of players that choose a path containing the edge $e$ in the outcome $(P_i, P_{-i})$. Then both sides of (13) are equal to

$$\sum_{e \in \tilde{P}_i \setminus P_i} \frac{c_e}{f_e + 1} - \sum_{e \in P_i \setminus \tilde{P}_i} \frac{c_e}{f_e}.$$

□

In the game theory literature, equation (13) is often taken as the definition of a potential function in the context of finite games. See Monderer and Shapley [24] for a fairly general treatment of potential functions for finite games.

While simple, Lemma 4.4 has a number of non-trivial consequences. First, Nash equilibria of a Shapley network design game are the local minima of the corresponding potential function. Formally, two outcomes of a Shapley network design game are *neighbors* if they differ in at most one component, and an outcome is a *local minimum* of $\Phi_{ND}$ if it has no neighbor with strictly smaller potential function value.

**Corollary 4.5** ([1], [30]). *An outcome of a Shapley network design game is a Nash equilibrium if and only if it is a local minimum of the corresponding potential function $\Phi_{ND}$.*

*Proof.* Immediate from the definitions and Lemma 4.4. □

Since every Shapley network design game has a finite number of outcomes, its corresponding potential function has a global (and hence local) minimum.

**Corollary 4.6** ([1], [30]). *In every Shapley network design game, there is at least one* (*pure-strategy*) *Nash equilibrium.*

We note in passing that several related classes of network games do not always have pure-strategy Nash equilibria [2], [6], [13], [30].

A stronger version of Corollary 4.6 also holds. In a finite game, *better-response dynamics* refers to the following process: start with an arbitrary initial outcome; if the current outcome is not a Nash equilibrium, pick an arbitrary player that can decrease its cost by switching strategies, update its strategy to an arbitrary superior one, and repeat. Better-response dynamics terminate if and only if a Nash equilibrium is reached. Even in extremely simple two-player games, better-response dynamics need not terminate (e.g., in "rock-paper-scissors"). On the other hand, the potential function (12) ensures that such dynamics always converge in Shapley network design games.

**Corollary 4.7** ([1], [30]). *In every Shapley network design game, better-response dynamics always converges to a Nash equilibrium in a finite number of iterations.*

*Proof.* By Lemma 4.4, every iteration of better-response dynamics strictly decreases the value of the potential function $\Phi_{ND}$. Better-response dynamics therefore cannot visit an outcome more than once and eventually terminates, necessarily at a Nash equilibrium. □

Corollary 4.7 does not address the number of iterations required to reach a Nash equilibrium; see [1], [12] for further study of this issue.

Finally, the potential function (12) has direct consequences for the price of stability in Shapley network design games. Comparing the definitions of cost (11) and potential function value (12) of such a game, we have

$$c(P_1, \ldots, P_k) \leq \Phi_{ND}(P_1, \ldots, P_k) \leq \mathcal{H}_k \cdot c(P_1, \ldots, P_k) \tag{14}$$

for every outcome $(P_1, \ldots, P_k)$. As a result, a global minimum of the potential function $\Phi_{ND}$ of a Shapley network design game is both a Nash equilibrium (by Corollary 4.6) and has cost at most $\mathcal{H}_k$ times that of optimal (by the argument in the proof of Theorem 2.10). This gives the following theorem

**Theorem 4.8** ([1]). *In every k-player Shapley network design game, the price of stability is at most* $\mathcal{H}_k$.

A similar argument shows that the bound of $\mathcal{H}_k$ in Theorem 4.8 applies to every Nash equilibrium reachable from an optimal solution via better-response dynamics. The bound also carries over to numerous extensions of Shapley network design games; see [1] for details.

Example 4.3 shows that the bound in Theorem 4.8 is tight for every $k \geq 1$. Thus, unlike for selfish routing networks and resource allocation games, a direct application

of a potential function argument yields an optimal upper bound on the inefficiency of equilibria.

The upper bound in Theorem 4.8 is not optimal for some important special cases of Shapley network design games, however. For example, suppose we insist that the underlying network $G$ is undirected. There is no known analogue of Example 4.3 for undirected Shapley network design games – the best lower bound known on the price of stability in such games is 2. On the other hand, it is not clear how to significantly improve the $\mathcal{H}_k$ bound in Theorem 4.8 for undirected networks.

**Open Question 2.** Determine the largest-possible price of stability in undirected Shapley network design games.

## 5. Conclusion

This survey has discussed three natural types of games: selfish routing networks, resource allocation games, and Shapley network design games. These classes of games differ from each other, both conceptually and technically, in a number of ways. Despite this, the worst-case inefficiency of selfish behavior is fairly well understood in all of these models, and in each case can be determined using a potential function characterization of equilibria.

While the entire field of algorithmic game theory is still in a relatively nascent stage, several broad research agendas are emerging. For the problem of quantifying the inefficiency of noncooperative equilibria, a central research issue is to understand characteristics of games that guarantee approximately optimal equilibria, and to develop flexible mathematical techniques for proving such guarantees. While many research accomplishments from the past few years have improved our understanding of these intertwined goals, there is clearly much left to be done. Perhaps the current state of the art in bounding the inefficiency of equilibria can be compared to the field of approximation algorithms circa twenty-five years ago, when the most fundamental problems and the most powerful algorithmic techniques (such as linear programming) were only beginning to crystallize. Motivated by this analogy, we conclude with the following question: will potential functions be as ubiquitous in bounds on the inefficiency of equilibria as linear programming is in bounds on the performance of approximation algorithms?

**Open Question 3.** We have seen that a potential function characterization of equilibria leads a bound on the inefficiency of equilibria. Under what conditions and to what extent does a converse hold? When does a bound on the inefficiency of the equilibria of a game imply the existence of some form of a potential function for the game?

## References


[1] Anshelevich, E., Dasgupta, A., Kleinberg, J., Tardos, É., Wexler, T., and Roughgarden, T., The price of stability for network design with fair cost allocation. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, IEEE Computer Society Press, Los Alamitos, CA, 2004, 295–304.

[2] Anshelevich, E., Dasgupta, A., Tardos, É., and Wexler, T., Near-optimal network design with selfish agents. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2003, 511–520.

[3] Awerbuch, B., Azar, Y., and Epstein, E., The price of routing unsplittable flow. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 57–66.

[4] Beckmann, M. J., McGuire, C. B., and Winsten, C. B. *Studies in the Economics of Transportation*. Yale University Press, 1956.

[5] Bertsekas, D. P., Nedic, A., and Ozdaglar, A. E., *Convex Analysis and Optimization*. Athena Scientific, 2003.

[6] Chen, H., and Roughgarden, T., Network design with weighted players. Submitted, 2005.

[7] Correa, J. R., Schulz, A. S., and Stier-Moses, N. E., Selfish routing in capacitated networks. *Math. Oper. Res.* **29** (4) (2004), 961–976.

[8] Correa, J. R., Schulz, A. S., and Stier-Moses, N. E., On the inefficiency of equilibria in congestion games. In *Integer Programming and Combinatorial Optimization*, Lecture Notes in Comput. Sci. 3509, Springer-Verlag, Berlin 2005, 167–181.

[9] Cramton, P., Shoham, Y., and Steinberg, R., *Combinatorial Auctions*. MIT Press, 2006.

[10] Doyle, P. G., and Snell, J. L., *Random Walks and Electrical Networks*. Mathematical Association of America, 1984.

[11] Dubey, P., Inefficiency of Nash equilibria. *Math. Oper. Res.* **11** (1) (1986), 1–8.

[12] Fabrikant, A., Papadimitriou, C. H., and Talwar, K., The complexity of pure Nash equilibria. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2004, 604–612.

[13] Fotakis, D., Kontogiannis, S. C., and Spirakis, P. G., Selfish unsplittable flows. *Theoret. Comput. Sci.* **348** (2–3) (2005), 226–239.

[14] Friedman, E. J., and Shenker, S., Learning and implementation on the Internet. Working paper, 1997.

[15] Garey, M. R., and Johnson, D. S., *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[16] Hajek, B., and Gopalakrishnan, G., Do greedy autonomous systems make for a sensible Internet? Presentation at the Conference on Stochastic Networks, Stanford University, June 2002 (Cited in [20]).

[17] Hartline, J. D., Optimization in the Private Value Model: Competitive Analysis Applied to Auction Design. PhD thesis, University of Washington, 2003.

[18] Jain, K., A polynomial time algorithm for computing the Arrow-Debreu market equilibrium for linear utilities. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, IEEE Computer Society Press, Los Alamitos, CA, 2004, 286–294.

[19] Johari, R., Efficiency Loss in Market Mechanisms for Resource Allocation. PhD thesis, MIT, 2004.

[20] Johari, R., and Tsitsiklis, J. N., Efficiency loss in a network resource allocation game. *Math. Oper. Res.* **29** (3) (2004), 407–435.

[21] Kelly, F. P., Network routing. *Philos. Trans. Roy. Soc. London Ser. A* **337** (3) (1991), 343–367.

[22] Kelly, F. P., Charging and rate control for elastic traffic. *European Transactions on Telecommunications* **8** (1) (1997), 33–37.

[23] Koutsoupias, E., and Papadimitriou, C. H., Worst-case equilibria. In *STACS 99*, Lecture Notes in Comput. Sci. 1563, Springer-Verlag, Berlin 1999, 404–413.

[24] Monderer, D., and Shapley, L. S., Potential games. *Games Econom. Behav.* **14** (1) (1996), 124–143.

[25] Nash, J. F., Equilibrium points in $N$-person games. *Proc. National Academy of Science* **36** (1) (1959), 48–49.

[26] Papadimitriou, C. H., Algorithms, games, and the Internet. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, ACM Press, New York 2001, 749–753.

[27] Pigou, A. C., *The Economics of Welfare*. Macmillan, 1920.

[28] Rapoport, A., and Chammah, A. M., *Prisoner's Dilemma*. University of Michigan Press, 1965.

[29] Rosenthal, R. W., A class of games possessing pure-strategy Nash equilibria. *Internat. J. Game Theory* **2** (1) (1973), 65–67.

[30] Rosenthal, R. W., The network equilibrium problem in integers. *Networks* **3** (1) (1973), 53–59.

[31] Roughgarden, T., The price of anarchy is independent of the network topology. *J. Comput. System Sci.* **67** (2) (2003), 341–364.

[32] Roughgarden, T., *Selfish Routing and the Price of Anarchy*. MIT Press, 2005.

[33] Roughgarden, T., Selfish routing and the price of anarchy. *OPTIMA* **71** (2006), to appear.

[34] Roughgarden, T., and Tardos, É., How bad is selfish routing? *J. ACM* **49** (2) (2002), 236–259.

[35] Schmeidler, D., Equilibrium points of nonatomic games. *J. Statist. Phys.* **7** (4) (1973), 295–300.

[36] Smith, M. J., The existence, uniqueness and stability of traffic equilibria. *Transportation Res. Part B* **13** (4) (1979), 295–304.

[37] Tardos, É., CS684 course notes. Cornell University, 2004.

[38] Wardrop, J. G., Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Pt. II*, volume 1, 1952, 325–378.

Stanford University, Department of Computer Science, 462 Gates Building, 353 Serra Mall, Stanford, CA 94305, U.S.A.

E-mail: tim@cs.stanford.edu

# Sublinear time algorithms

Ronitt Rubinfeld*

**Abstract.** Sublinear time algorithms represent a new paradigm in computing, where an algorithm must give some sort of an answer after inspecting only a very small portion of the input. We discuss the sorts of answers that one might be able to achieve in this new setting.

**Mathematics Subject Classification (2000).** Primary 68Q25; Secondary 68W20, 68W25.

**Keywords.** Sublinear time algorithms, property testing.

## 1. Introduction

The goal of algorithmic research is to design efficient algorithms, where efficiency is typically measured as a function of the length of the input. For instance, the elementary school algorithm for multiplying two $n$ digit integers takes roughly $n^2$ steps, while more sophisticated algorithms have been devised which run in less than $n \log^2 n$ steps. It is still not known whether a linear time algorithm is achievable for integer multiplication. Obviously any algorithm for this task, as for any other nontrivial task, would need to take at least linear time in $n$, since this is what it would take to read the entire input and write the output. Thus, showing the existence of a linear time algorithm for a problem was traditionally considered to be the gold standard of achievement.

Nevertheless, due to the recent tremendous increase in computational power that is inundating us with a multitude of data, we are now encountering a paradigm shift from traditional computational models. The scale of these data sets, coupled with the typical situation in which there is very little time to perform our computations, raises the issue of whether there is time to consider any more than a miniscule fraction of the data in our computations? Analogous to the reasoning that we used for multiplication, for most natural problems, an algorithm which runs in sublinear time must necessarily use randomization and must give an answer which is in some sense imprecise. Nevertheless, there are many situations in which a fast approximate solution is more useful than a slower exact solution.

A first example of sublinear time computation that comes to mind is the classical result from the theory of sampling that one can, in time independent of the size of the data, determine a good estimate of the average value of a list of numbers of bounded

*supported by NSF grant 012702-001

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

magnitude. But what about more interesting algorithmic questions? For example, given access to all transcripts of trades in the stock exchange, can we determine whether there is a trend change? This is easily detectable after a careful scan of the entire transcript, but by the time the scan is performed, it might be too late to make use of the information. However, it might be feasible to construct much faster algorithms based on random sampling. The recently emerging theory of sublinear time algorithms addresses questions of precisely this nature for problems in various domains.

This paper will describe a number of problems that can be solved in sublinear time, using different types of approximations.

**Outline of the paper.** We begin by giving a motivating example of a sublinear time algorithm in Section 2. In Section 3 we formalize the definitions of approximations that sublinear time algorithms are able to achieve. We then describe various examples of sublinear time algorithms.

## 2. A simple example: monotonicity of a list

Let us begin with a simple example, which will motivate our subsequent definitions of the types of approximations that we will be interested in achieving. A list of integers $\vec{x} = x_1, \ldots, x_n$, is *monotone* (increasing) if $x_i \leq x_j$ for all $1 \leq i < j \leq n$. Given input $\vec{x}$, the task is to determine whether or not $\vec{x}$ is monotone.

In order to construct an algorithm that runs in sublinear time for determining whether $\vec{x}$ is monotone, we first need to make our model of computation precise. For example, if our algorithm must scan $x_1, \ldots, x_{i-1}$ in order to reach $x_i$, then there is no hope for the existence of a sublinear time algorithm. However, it is often natural to assume that our algorithms have *query* (also called *oracle*) access to the input. That is, they can access $x_i$ in one step for any $1 \leq i \leq n$.

Even with this model of computation, it is clear that finding a sublinear time algorithm for the above task is impossible, since any algorithm that does not look at some $x_j$ could be fooled by an input for which all the $x_i$'s are in monotone order for $i \neq j$. Thus, we can only hope to solve an approximate version of this problem, but what is a meaningful notion of an approximation?

One natural approximate version is defined as follows: Say that $x_1, \ldots, x_n$ is *$\varepsilon$-close* to monotone if by changing at most $\varepsilon n$ of the values of the $x_i$'s one can transform $x_1, \ldots, x_n$ into a monotone list. Then, a *property tester for monotonicity* is a randomized algorithm that on input $\vec{x}, \varepsilon$, must output "pass" if $x_1, \ldots, x_n$ is monotone, and "fail" if $x_1, \ldots, x_n$ is not $\varepsilon$-close to monotone. The algorithm is allowed to err with probability at most $1/3$. However, once an algorithm with error probability at most $1/3$ is achieved, it is easy to see that for any $\beta$, a probability of error of at most $\beta$ can be achieved by repeating the algorithm $O\left(\log \frac{1}{\beta}\right)$ times and taking the majority answer. Note that if $x_1, \ldots, x_n$ is, say, $\varepsilon/2$-close to monotone,

the property testing algorithm is allowed to output "pass" or "fail". Indeed, in this case, since the list is close to monotone, it may not be too harmful to pass it. On the other hand, since it is not actually monotone, it is also not a big injustice to fail it.

How do we construct a property tester for monotonicity? On first thought, one might try picking random indices $i$, $j$ and performing tests of the form "is $x_i \leq x_j$?" or "is $x_i \leq x_{i+1}$?". However, these tests do not work very well. It is easy to construct examples showing that there are lists that are not even $\left(1 - \frac{1}{\sqrt{n}}\right)$-close to monotone, yet pass such tests with probability at least $1 - \frac{1}{n^{1/4}}$. This means that at least $n^{1/4}$ such tests must be performed if one is to find a reason to output "fail". Though this does not rule out the possibility of a sublinear time property tester, we will see that one can do much better. In the following, we describe an $O(\log n)$ time algorithm from the work of Ergün et. al. [17] which tests if $\vec{x}$ has a long monotone increasing subsequence. Note that the problem is known to require $\Omega(\log n)$ queries [17], [20].

Let $c$ be a constant that is set appropriately. For simplicity, let us assume that the elements in $\vec{x}$ are distinct. The last assumption is without loss of generality, since one can append the index of an item to the least significant bits of its value in order to break ties.

1. Let $\ell = c/\varepsilon$. Choose indices $i_1, \ldots, i_\ell$ uniformly from $[n]$.
2. For each such chosen index $i_j$, assume the list is monotone and perform a binary search in $\vec{x}$ as if to determine whether $x_{i_j}$ is present in $\vec{x}$ or not.
3. Output "fail" if the binary search fails to find any $x_{i_j}$ in location $i_j$ or finds a pair of out-of-order elements along the search path. Output "pass" if all the $\ell$ binary searches succeed.

The running time of the algorithm is $O((1/\varepsilon)\log n)$. Moreover, if $\vec{x}$ is monotone, then the algorithm will always output "pass" as each of the binary searches will succeed. To show that if $\vec{x}$ is not $\varepsilon$-close to monotone, then the algorithm will output "fail" with probability at least $2/3$, we show the contrapositive. Namely, assume that the input is such that the algorithm outputs "pass" with probability at least $1/3$. To see that $\vec{x}$ has a long increasing subsequence, let $G \subseteq [n]$ denote the set of indices for which the binary search would succeed, i.e., $i \in G$ if and only if $x_i$ can be found by a binary search on $\vec{x}$ that sees no pair of out-of-order elements along the search path. The constant $c$ can be chosen such that if $|G| < (1 - \varepsilon)n$, then the algorithm would pick some $i_j \notin G$ with probability at least $1/3$, causing it to output "fail". Thus, since the algorithm outputs "pass" with probability at least $1/3$, we know that $|G| \geq (1 - \varepsilon)n$. We now argue that the restriction of $\vec{x}$ to the indices in $G$ is an increasing subsequence: Let $i, j \in G$ and $i < j$. Let $k$ be the least common ancestor index where the binary searches for $x_i$ and $x_j$ diverge. Then $x_i < x_k$ and $x_k < x_j$, which implies $x_i < x_j$. Finally, if $\vec{x}$ has an increasing subsequence of size at least $(1 - \varepsilon)n$ then it is easy to see that $\vec{x}$ is $\varepsilon$-close to monotone. Thus we have the following theorem:

**Theorem 2.1** ([17]). *There is an algorithm that, given a sequence $\vec{x} = x_1, \ldots, x_n$ and an $\varepsilon > 0$, runs in $O((1/\varepsilon)\log n)$ time and outputs* (1) "pass"*, if $\vec{x}$ is monotone*

*and* (2) "fail"*, with probability* $2/3$*, if* $\vec{x}$ *does not have an increasing subsequence of length at least* $(1-\varepsilon)n$ (*in particular, if* $\vec{x}$ *is* $\varepsilon$*-far from monotone*).

## 3. What do we mean by an "approximation"?

Now that we have developed some intuition, we present our model and definitions in more detail: We are interested in computing some function $f$ on input $x$ without reading all of $x$. This is an impossible task in general, since a change to a single bit of $x$ could alter the value of $f$. When $f$ is the characteristic function of a property, i.e., $f(x) = 1$ if $x$ has the property and $f(x) = 0$ otherwise, the following notion of approximation has emerged: Given an input, a *property tester* tries to distinguish whether the input has the property from the case where the input is not even close to having the property. We first formalize what it means to be close.

**Definition 3.1.** An input $x$, represented as a function $x\colon \mathcal{D} \to \mathcal{R}$, is $\varepsilon$*-close* to satisfying property $P$ if there is some $y$ satisfying $P$ such that $x$ and $y$ differ on at most $\varepsilon|\mathcal{D}|$ places in their representation. Otherwise, $x$ is said to be $\varepsilon$*-far* from satisfying $P$.

In the monotonicity example of the previous section, $\mathcal{D} = [n]$ and $x(i)$ returns the $i^{\text{th}}$ element of the list.

We now formalize what it means for an algorithm to test a property. As in the previous section, we assume in our model of computation that algorithms have query access to the input.

**Definition 3.2.** Let $P$ be a property. On input $x$ of size $n = |\mathcal{D}|$ and $\varepsilon$, a *property tester* for $P$ must satisfy the following:

- If $x$ satisfies property $P$, the tester must output "pass" with probability at least $2/3$.
- If $x$ is $\varepsilon$-far from satisfying $P$, the tester must output "fail" with probability at least $2/3$.

The probability of error may depend only on the coin tosses of the algorithm and not on any assumptions of the input distribution. The number of queries made by the property tester $q = q(\varepsilon, n)$ is referred to as the query complexity of the property tester. We say that a property tester for $P$ has 1*-sided error* if it outputs "pass" with probability 1 when $x$ satisfies $P$. If the query complexity is independent of $n$, then we say that the property is *easily testable*.

Note that if $x$ does not satisfy $P$ but $x$ is also not $\varepsilon$-far from satisfying $P$, then the output of the property tester can be either "pass" or "fail". We have already seen that it is this gap which allows property testers to be so efficient.

The probability that the property tester errs is arbitrarily set to $1/3$ and may alternatively be defined to be any constant less than 1/2. It is then easy to see that for any $\beta$, a probability of error of at most $\beta$ can be achieved by repeating the algorithm $O\big(\log \frac{1}{\beta}\big)$ times and taking the majority answer.

Property testing was first defined by Rubinfeld and Sudan [37] in the context of program testing. Goldreich, Goldwasser, Ron [23] refined and generalized the definition. Various more general definitions are given in several works, including [17], [25], [31], which mostly differ in terms of the generality of the distance function and natural generalizations as to when the tester should accept and reject.

## 4. Algebraic problems

In this section, we consider property testing algorithms for problems that are algebraic in nature. We begin with the problem of testing whether a function is a homomorphism. We then show that the ideas used to construct property testers for homomorphisms extend to other properties with similar underlying structure.

**4.1. Homomorphism testing.** We begin with an example that was originally motivated by applications in program testing [14] and was later used in the construction of Probabilistically Checkable Proof systems [6]. Suppose you are given oracle access to a function $f : \mathcal{D} \to \mathcal{R}$, that is, you may query the oracle on any input $x \in \mathcal{D}$ and it will reply with $f(x)$. Is $f$ a homomorphism?

In order to determine the answer exactly, it is clear that you need to query $f$ on the entire domain $\mathcal{D}$. However, consider the property testing version of the problem, for which on input $\varepsilon$, the property tester should output "pass" with probability at least $2/3$ if $f$ is a homomorphism and "fail" with probability at least $2/3$ if $f$ is $\varepsilon$-far from a homomorphism (that is, there is no homomorphism $g$ such that $f$ and $g$ agree on at least $(1-\varepsilon)|\mathcal{D}|$ inputs). In order to construct such a property tester, a natural idea would be to test that the function satisfies certain relationships that all homomorphisms satisfy. We next describe two such relationships and discuss their usefulness in constructing property testers.

**Two characterizations of homomorphisms over $\mathbb{Z}_q$.** Consider the case when $f$ is over the domain and range $\mathcal{D} = \mathcal{R} = \mathbb{Z}_q$ for large integer $q$. The set of homomorphisms over $\mathbb{Z}_p$ can be characterized as the set of functions which satisfy $f(x+1) - f(x) = f(1)$ for all $x$. This suggests that a property tester might test that $f(x+1) - f(x) = f(1)$ for most $x$. However, it is easy to see that there are functions $f$ which are very far from any homomorphism, but would pass such a test with overwhelmingly high probability. For example, $g(x) = x \bmod \lceil\sqrt{q}\rceil$ satisfies $g(x+1) - g(x) = g(1)$ for $1 - \frac{1}{\sqrt{q}}$ fraction of the $x \in \mathbb{Z}_q$ but $g(x)$ is $\big(1 - \frac{1}{\sqrt{q}}\big)$-far from a homomorphism.

The set of homomorphisms over $\mathcal{D}$ can alternatively be characterized as the set of functions which satisfy $f(x) + f(y) = f(x + y)$ for all $x, y$. This suggests that one might test that $f(x) + f(y) = f(x + y)$ for most $x, y$. It might be worrisome to note that when $q = 3n$, the function $h(x)$ defined by $h(x) = 0$ if $x \equiv 0 \bmod 3$, $h(x) = 1$ if $x \equiv 1 \bmod 3$ and $h(x) = 3n - 1$ if $x \equiv -1 \bmod 3$ passes the above test for 7/9 fraction of the choices of pairs $x, y \in \mathcal{D}$ and that $h(x)$ is 2/3-far from a homomorphism [16]. However, here the situation is much different: one can show that for any $\delta < 2/9$, if $f(x) + f(y) = f(x + y)$ for at least $1 - \delta$ fraction of the choices of $x, y \in \mathcal{D}$, then there is some homomorphism $g$, such that $f(x) = g(x)$ on at least $1 - \delta/2$ fraction of the $x \in \mathcal{D}$ [13].

Once one has established such a theorem, then one can construct a property tester based on this characterization by sampling $O(1/\varepsilon)$ pairs $x, y$ and ensuring that each pair in the sample satisfies $f(x) + f(y) = f(x + y)$. This property tester clearly passes all homomorphisms. On the other hand, if $f$ is $\varepsilon$-far from a homomorphism then the above statement guarantees that at least $2\varepsilon$ fraction of the choices of $x, y$ pairs do not satisfy $f(x) + f(y) = f(x + y)$, and the property tester is likely to fail.

In both cases, homomorphisms are characterized by a collection of *local* constraints, where by local, we mean that few function values are related within each constraint. What is the difference between the first and the second characterization of a homomorphism that makes the former lead to a bad test and the latter to a much better test? In [37] (see also [36]), the notion of a *robust characterization* was introduced to allow one to quantify the usefulness of a characterization in constructing a property test. Loosely, a robust characterization is one in which the "for all" quantifier can be replaced by a "for most" quantifier while still characterizing essentially the same functions. That is, for a given $\varepsilon, \delta$, a characterization is $(\varepsilon, \delta)$-*robust* if for any function $f$ that satisfies at least $1 - \delta$ fraction of the constraints, $f$ must be $\varepsilon$-close to some function $g$ that satisfies all of the constraints and is thus a solution of the "for all" characterization. As we saw above, once we have an $(\varepsilon, \delta)$-robust characterization for a property, it is a trivial matter to construct a property tester for the property. We are interested in the relationship between $\varepsilon$ and $\delta$ as well as the range of $\delta$ for which the property is $(\varepsilon, \delta)$-robust, since the value of $\delta$ directly influences the running time of the property tester.

**Homomorphism testing, a history.** Let $G, H$ be two finite groups. For an arbitrary map $f: G \to H$, define $\delta$, the probability of group law failure, by

$$1 - \delta = \Pr_{x,y} [f(x) + f(y) = f(x + y)] .$$

Define $\varepsilon$ such that $\varepsilon$ is the minimum $\tau$ for which $f$ is $\tau$-close to a homomorphism. We will be interested in the relationship between $\varepsilon$ and $\delta$.

Blum, Luby and Rubinfeld [14], considered this question and showed that over cyclic groups, there is a constant $\delta_0$, such that if $\delta \leq \delta_0$, then the one can upper bound $\varepsilon$ in terms of a function of $\delta$ that is independent of $|G|$. This yields a homomorphism

tester with query complexity that depends (polynomially) on $1/\varepsilon$, but is independent of $|G|$, and therefore shows that the property of being a homomorphism is easily testable. The final version of [14] contains an improved argument due to Coppersmith [16], which applies to all Abelian groups, shows that $\delta_0 < 2/9$ suffices, and that $\varepsilon$ is upper bounded by the smaller root of $x(1-x) = \delta$ (yielding a homomorphism tester with query complexity linear in $1/\varepsilon$). Furthermore, the bound on $\delta_0$ was shown to be tight for general groups [16]. In [13], it was shown that for general (non-Abelian) groups, for $\delta_0 < 2/9$, then $f$ is $\varepsilon$-close to a homomorphism where $\varepsilon = (3 - \sqrt{9 - 24\delta})/12 \leq \delta/2$ is the smaller root of $3x - 6x^2 = \delta$. The condition on $\delta$, and the bound on $\varepsilon$ as a function of $\delta$, are shown to be tight, and the latter improves that of [14], [16]. Though $\delta_0 < 2/9$ is optimal over general Abelian groups, using Fourier techniques, Bellare et. al. [12] have shown that for groups of the form $(\mathbb{Z}/2)^n$, $\delta_0 \leq 45/128$ suffices.

**A proof of a homomorphism test.** We describe the following proof, essentially due to Coppersmith [14], [16] of the robustness of the homomorphism characterization over Abelian groups. Though this is not the strongest known result, we include this proof to give a flavor of the types of arguments used to show robustness of algebraic properties.

**Theorem 4.1.** *Let $G$ be a finite Abelian group and $f : G \to G$. Let $\delta$ be such that*

$$1 - \delta = \Pr_{x,y}[f(x) + f(y) = f(x+y)].$$

*Then if $\delta < 2/9$, $f$ is $2\delta$-close to a homomorphism.*

*Proof.* Define $\phi(x) = \mathrm{maj}_{y \in G}(f(x+y) - f(y))$, that is, let $\phi(x)$ be the value that occurs with the highest probability when evaluating $f(x+y) - f(y)$ over random $y$ (breaking ties arbitrarily).

The theorem follows immediately from the following two claims showing that $\phi$ is a homomorphism and that $f$ and $\phi$ are $2\delta$-close.

**Claim 4.2.** $|\{y | f(y) = \phi(y)\}| \geq (1 - 2\delta)|G|$.

*Proof of Claim* 4.2. Let $B = \{x \in G : \Pr_y[f(x) \neq f(x+y) - f(y)] > 1/2]\}$. If $x \notin B$, then $\phi(x) = f(x)$. Thus, it suffices to bound $\frac{|B|}{|G|}$. If $x \in B$, then $\Pr_y[f(x) + f(y) \neq f(x+y)] > 1/2$. Thus $\delta = \Pr_{x,y}[f(x) \neq f(x+y) - f(y)] \geq \frac{|B|}{|G|} \cdot \frac{1}{2}$ or equivalently $\frac{|B|}{|G|} \leq 2\delta$. □

**Claim 4.3.** *If $\delta < 2/9$, then $\phi(x) + \phi(z) = \phi(x+z)$ for all $x$, $z$.*

*Proof of Claim* 4.3. Fix $x$, we first show that most pairs $y_1$, $y_2$ agree to vote for the same value of $\phi(x)$. Pick random $y_1, y_2 \in G$, and we have:

$$\begin{aligned}\Pr_{y_1,y_2}[f(x+y_1) - f(y_1) = f(x+y_2) - f(y_2)] &= \Pr_{y_1,y_2}[f(x+y_1) + f(y_2) \\ &= f(x+y_2) + f(y_1)].\end{aligned}$$

$x + y_1$ and $y_2$ are both uniformly distributed elements of $G$. Therefore we have $\Pr_{y_1,y_2}[f(x + y_1) + f(y_2) \neq f(x + y_1 + y_2)] = \delta < 2/9$. Similarly, we have $\Pr_{y_1,y_2}[f(x + y_2) + f(y_1) \neq f(x + y_1 + y_2)] = \delta < 2/9$. If neither of the above events happens, then $f(x + y_1) - f(y_1) = f(x + y_2) - f(y_2)$. Via a union bound we have that

$$\Pr_{y_1,y_2}[f(x + y_1) - f(y_1) = f(x + y_2) - f(y_2)] > 1 - 2\delta \geq 5/9.$$

It is straightforward to show that for any distribution in which the collision probability is at least $5/9$, the maximum probability element must have probability at least $2/3$. Thus,

$$\Pr_y[\phi(x) \neq f(x + y) - f(y)] < 1/3 \quad \text{for all } x \in G. \tag{1}$$

To show that for all $x, z \in G$, $\phi(x) + \phi(z) = \phi(x + z)$, fix $x$ and $z$. Then apply Equation (1) to $x$, $z$ and $x + z$ to get

$$\Pr_y[\phi(x) \neq f(x + (y - x)) - f(y - x)] < 1/3, \tag{2}$$

$$\Pr_y[\phi(z) \neq f(z + y) - f(y)] < 1/3, \tag{3}$$

$$\Pr_y[\phi(x + z) \neq f((x + z) + (y - x)) - f(y - x)] < 1/3. \tag{4}$$

Thus $\Pr_y[\phi(x) = f(y) - f(y - x)$ and $\phi(z) = f(z + y) - f(y)$ and $\phi(x + z) = f(z + y) - f(y - x)] > 0$, and so there exists a $y$ for which

$$\phi(x)+\phi(z) = (f(y)-f(y-x))+(f(z+y)-f(y)) = f(z+y)-f(y-x) = \phi(x+z).$$

The above equality holds for every $x, z \in G$, showing that $\phi$ is a homomorphism and completing the proof of Claim 4.3. □

□

**A word about self-correcting.** In the proof, we note that $\phi$ is defined so that it is the "self-correction" of $f$. Observe that there is a simple randomized algorithm that computes $\phi(x)$ given oracle access to $f$: pick $c \log 1/\beta$ values $y$, compute $f(x + y) - f(y)$ and output the value that you see most often. If $f$ is $\frac{1}{8}$-close to a homomorphism $\phi$, then since both $y$ and $x + y$ are uniformly distributed, we have that for at least $3/4$ of the choices of $y$, $\phi(x + y) = f(x + y)$ *and* $\phi(y) = f(y)$, in which case $f(x + y) - f(y) = \phi(x)$. Thus it is easy to show that there is a constant $c$ such that if $f$ is $\frac{1}{8}$-close to a homomorphism $\phi$, then the above algorithm will output $\phi(x)$ with probability at least $1 - \beta$.

**4.2. Other algebraic functions.** It is natural to wonder what other classes of functions have robust characterizations as in the case of homomorphisms? There are many other classes of functions that are defined via characterizations that are local. The field of functional equations is concerned with the prototypical problem of characterizing the set of functions that satisfy a given set of properties (or functional equations). For

example, the class of functions of the form $f(x) = \tan Ax$ are characterized by the functional equation

$$f(x+y) = \frac{f(x)+f(y)}{1-f(x)f(y)} \quad \text{for all } x, y.$$

D'Alembert's equation

$$f(x+y)+f(x-y) = 2f(x)f(y) \quad \text{for all } x, y$$

characterizes the functions $0, \cos Ax, \cosh Ax$. Multivariate polynomials of total degree $d$ over $\mathbb{Z}_p$ for $p > md$ can be characterized by the equation

$$\sum_{i=0}^{d+1} \alpha_i f(\hat{x} + i\hat{h}) = 0 \quad \text{for all } \hat{x}, \hat{h} \in Z_p^m,$$

where $\alpha_i = (-1)^{i+1}\binom{d+1}{i}$. All of the above characterizations are known to be $(\varepsilon, \delta)$-robust for $\varepsilon$ and $\delta$ independent of the domain size (though for the case of polynomials, there is a polynomial dependence on the total degree $d$) thus showing that the corresponding properties are easily testable [36], [37]. A long series of works have given increasingly robust characterizations of functions that are low total degree polynomials (cf. [6], [32], [7], [34], [3], [29], [27]).

We note that all of these results can be extended to apply over domains that are subsets of infinite cyclic groups. They can further be extended to the case of computation with finite precision, which requires that one address the stability of functional equations [18], [30].

**Convolutions of distributions.** We now turn to a seemingly unrelated question about distributions that are close to their self-convolutions: Let $A = \{a_g \mid g \in G\}$ be a distribution on group $G$. The convolution of distributions $A$, $B$ is

$$C = A * B, \quad c_x = \sum_{\substack{y,z \in G \\ yz = x}} a_y b_z.$$

Let $A'$ be the *self-convolution* of $A$, $A * A$, i.e. $a'_x = \sum_{y,z \in G; yz=x} a_y a_z$. It is known that $A = A'$ exactly when $A$ is the uniform distribution over a subgroup of $G$. Suppose we know that $A$ is close to $A'$, can we say anything about $A$ in this case? Suppose $\text{dist}(A, A') = \frac{1}{2}\sum_{x \in G} |a_x - a'_x| \leq \varepsilon$ for small enough $\varepsilon$. Then [13] show that $A$ must be close to the uniform distribution over a subgroup of $G$. More precisely, in [13] it is shown that for a distribution $A$ over a group $G$, if $\text{dist}(A, A') = \frac{1}{2}\sum_{x \in G} |a_x - a'_x| \leq \varepsilon \leq 0.0273$, then there is a subgroup $H$ of $G$ such that $\text{dist}(A, U_H) \leq 5\varepsilon$, where $U_H$ is the uniform distribution over $H$ [13]. On the other hand, in [13] there is an example of a distribution $A$ such that $\text{dist}(A, A * A) \approx .1504$, but $A$ is not close to uniform on any subgroup of the domain.

A weaker version of this result, was used to prove a preliminary version of the homomorphism testing result in [14]. To give a hint of why one might consider the question on convolutions of distributions when investigating homomorphism testing, consider the distribution $A_f$ achieved by picking $x$ uniformly from $G$ and outputting $f(x)$. It is easy to see that the error probability $\delta$ in the homomorphism test is at least $\text{dist}(A_f, A_f * A_f)$. The other, more useful, direction is less obvious. In [13] it is shown that this question on distributions is "equivalent" in difficulty to homomorphism testing:

**Theorem 4.4.** *Let $G$, $H$ be finite groups. Assume that there is a parameter $\beta_0$ and function $\phi$ such that the following holds:*

> *For all distributions $A$ over group $G$, if $\text{dist}(A * A, A) \leq \beta \leq \beta_0$ then $A$ is $\phi(\beta)$-close to uniform over a subgroup of $G$.*

*Then, for any $f : G \to H$ and $\delta < \beta_0$ such that $1 - \delta = \Pr[f(x) * f(y) = f(x * y)]$, and $\phi(\delta) \leq 1/2$, we have that $f$ is $\phi(\delta)$-close to a homomorphism.*

## 5. Combinatorial objects

In 1996, Goldreich, Goldwasser and Ron [23] focused attention on the problem of testing various properties of graphs and other combinatorial objects. Their work introduced what is now referred to as the *dense graph model* of property testing. In this model, a graph on $n$ nodes is represented via an $n \times n$ adjacency matrix, where the $(i, j)$th entry of the matrix contains a 1 if the edge $(i, j)$ is present in the graph and a 0 otherwise. Two graphs $G$ and $H$ are $\varepsilon$-close if at most $\varepsilon n^2$ edges need to be modified (inserted or deleted) to turn $G$ into $H$. In [23], several graph properties were shown to be easily testable. In fact, as we shall soon see, the question of which graph properties are easily testable has led to a series of intriguing results.

One property that [23] consider is that of $k$-colorability – is it possible to assign one of $k$ colors to each of the nodes so that no pair of nodes that have an edge between them are assigned the same color? The property of $k$-colorability is NP-complete to determine – meaning, that though we know how to verify that a certain coloring is a valid $k$-coloring, we have no idea how to determine whether a graph has a $k$ coloring in time polynomial in the size of the graph. Somewhat surprisingly, $k$-colorability is easily testable, so we can distinguish $k$-colorable graphs from those that are $\varepsilon$-far from $k$-colorable in constant time. Thus we see that the efficiency of a property tester is not directly related to the complexity of deciding the property exactly.

Though the proof of correctness of the property tester for $k$-colorability is involved, the algorithm used to conduct the property test is easy to describe: It simply picks a constant sized random sample of the vertices, queries all the edges among this random sample and then outputs "pass" or "fail", according to whether the sample is

$k$-colorable. Since the sample is of constant size, the determination of whether the sample is $k$-colorable can be made in constant time.

Such algorithms that (1) pick a constant sized random sample of the vertices, (2) query all the edges among this random sample, and then (3) output "pass" or "fail" based on whether the subgraph has the property or not, are referred to as "natural algorithms". Modulo a technicality about how the final output decision is made, Goldreich and Trevisan [26] essentially show that for any graph property that is easily testable, the natural algorithm gives a property tester. Thus, *all* easily testable graph properties provably have easy-to-describe algorithms.

The work of [23] sparked a flurry of results in the dense graph model. A very interesting line of work was initiated in the work of Alon, Fischer, Krivelevich and Szegedy [2], in which they use the Szemerédi Regularity Lemma to show that the property of a graph being $H$-free (that is, the graph does not contain any copy of $H$ as a subgraph) is easily testable for any constant sized graph $H$.

Very recently, the above line of work culminated in the following amazing result: Alon and Shapira [5] have shown that one can *completely* characterize the classes of graph properties that are easily testable with 1-sided error in the dense graph model. Before describing their result, we make two definitions. A graph property $P$ is *hereditary* if it is closed under the removal of vertices (but not necessarily under the removal of edges). A graph property $P$ is *semi-hereditary* if there is a hereditary graph property $H$ such that (1) any graph satisfying $P$ also satisfies $H$ and (2) for any $\varepsilon > 0$, there is an $M(\varepsilon)$, such that any graph $G$ of size at least $M(\varepsilon)$ which is $\varepsilon$-far from satisfying $P$ does not satisfy $H$. The result of Alon and Shapira is then the following:

**Theorem 5.1.** *A graph property $P$ is easily testable with one-sided error if and only if $P$ is semi-hereditary.*

Hereditary graph properties include all monotone graph properties (including $k$-colorability and $H$-freeness), as well as other interesting non-monotone graph properties such as being a perfect, chordal, or interval graph. The techniques used by Alon and Shapira are quite involved, and are based on developing new variants of the Szemerédi Regularity Lemma. Previously in the literature, the "Regularity Lemma type" arguments were used to develop testers for graph properties that were characterized by a finite set of forbidden subgraphs. Here the set of forbidden subgraphs may be infinite, and they are forbidden as *induced* subgraphs.

Several interesting questions regarding easily testable graph properties remain. For example, because of the use of the Szemerédi Regularity Lemma, the upper bounds given by the previously mentioned results have a dependence on $1/\varepsilon$ that is enormous. It would be interesting to characterize which problems have property testers whose dependence on $1/\varepsilon$ is polynomial (cf. [1]).

There are many interesting properties that are not easily testable, but do have sublinear time property testers. For example, the graph isomorphism problem asks whether two graphs are identical under relabeling of the nodes. In [21], it is shown

that the property testing problem requires $\Omega(n)$ queries and that there is a property tester for this problem which uses $O(n^{5/4} \operatorname{polylog} n)$ queries, which is sublinear in the input size $n^2$.

The area of property testing has been very active, with a number of property testers devised for other models of graphs as well as other combinatorial objects. The testability of a problem is very sensitive to the model in which it is being tested. In contrast to the dense graph model, where $k$-colorability is easily testable, it is known that there are no sublinear time property testers for the $k$-colorability problem in models suitable for testing sparse graphs [15]. Property testers have also been studied in models of general graphs, and threshold-like behaviors have been found for the complexity of the testing problems in terms of the average degree of the graph [28], [4].

Property testers for combinatorial properties of matrices, strings, metric spaces and geometric objects have been devised. We refer the reader to the excellent surveys of Goldreich [22], Ron [35] and Fischer [19].

## 6. Testing properties of distributions

In a wide variety of settings, data is most naturally viewed as coming from a probability distribution. In order to effectively make use of such data, one must understand various properties of the underlying probability distribution. Some of these properties are "local" in nature, for example focusing on whether or not a specific domain element appears with large probability. Other properties have a rather "global" feel in the sense that they are a property of the distribution as a whole and not of a small subset of the domain elements. Unlike the case for local properties, it makes sense to characterize a distribution in terms of some meaningful distance measure to the closest distribution that has the global property. This yields a somewhat different model than the property testing model in terms of the assumption on how the data is presented for here we do not assume that an explicit description of the distribution is given.

In the following, we assume that there is an underlying distribution from which the testing algorithm receives independent identically distributed (iid) samples. The complexity of the algorithm is measured in terms of the number of samples required in order to produce a meaningful answer (the sample complexity). As mentioned in the introduction, it is a classical result from the theory of sampling that one can, in time independent of the size of the data, determine a good estimate of the average value of a list of numbers of bounded magnitude. However, more recently, properties such as closeness between two distributions, closeness to an explicitly given distribution, independence, and high entropy, have been studied in this model [24], [10], [9]. For many properties, well-known statistical techniques, such as the $\chi^2$-test or the straightforward use of Chernoff bounds, have sample complexities that are at least linear in the size of the support of the underlying probability distribution. In contrast, there are algorithms whose sample complexity is sublinear in the size of the support

for various properties of distributions.

We mention one illustrative example: Given samples of a distribution $X$ on $[n]$, for example all the previous winners of the lottery, how can one tell whether $X$ is close to uniform? We will measure closeness in terms of the $L_2$ norm, i.e., letting $X(i)$ denote the probability that $X$ assigns to $i$,

$$\|X\|_2 = \sum_{i\in[n]} (X(i) - 1/n)^2.$$

Goldreich and Ron [24] note that since

$$\sum_{i\in[n]} (X(i) - 1/n)^2 = \sum_{i\in[n]} X(i)^2 - 1/n,$$

it is enough to estimate the collision probability. They then show that this can be done by considering only $O(\sqrt{n})$ samples and counting the number of pairs that are the same. By bounding the variance of their estimator, they obtain the following:

**Theorem 6.1** ([24]). *There is an algorithm that, given a distribution $X$ on $[n]$ via a generation oracle, approximates $\|X\|_2$ to within a multiplicative factor of $(1 \pm \varepsilon)$ using $O(\sqrt{n}/\varepsilon^2)$ samples, with constant probability.*

Such techniques are very useful for achieving sublinear time algorithms for testing whether distributions satisfy several other global properties. For example, for the properties of closeness of two arbitrary distributions [10], independence of a joint distribution [9], high entropy [8], and monotonicity of the probability density function (when the distribution is over a totally ordered domain) [11], the testing problem can be reduced to the problem of testing the near-uniformity of the distribution on various subdomains.

## 7. Some final comments

We have seen several contexts in which one can test properties in sublinear time. The study of sublinear time algorithms has led to a new understanding of many problems that had already been well-studied. Though we have mentioned only property testing problems in this survey, other, more traditional, types of approximations are achievable in sublinear time. Such algorithms have been used to design very fast approximation algorithms for graph problems and for string compressibility problems (cf. [23], [10], [33]). Some of these algorithms have even resulted in better linear time approximation algorithms than what was previously known.

Probabilistically Checkable Proof Systems (PCPs) can be thought of as a way to write down a proof so that another person can verify it by viewing only a constant number of locations (cf. [6]). PCPs can thus be viewed as a type of robust characterization and their verification is a sublinear algorithm. More interestingly, property

testers for homomorphisms and low degree polynomials are used as key ingredients in the construction of Probabilistically Checkable Proof Systems.

As we have seen, the study of sublinear algorithms gives a new perspective that has yielded insights to other areas of theoretical computer science. Much still remains to be understood about the scope of sublinear time algorithms, and we expect that this understanding will lead to further insights.

**Acknowledgements.** We thank Ran Canetti, Tali Kaufman, and Madhu Sudan for their extensive and helpful comments on this manuscript.

## References

[1] Alon, N., Testing subgraphs in large graphs. In *Proceedings of the 46th Symposium on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2001, 434–441.

[2] Alon, N., Fischer, E., Krivelevich, M., and Szegedy, M., Efficient testing of large graphs. *Combinatorica* **20** (2000), 451–476.

[3] Alon, N., Kaufman, T., Krivelevich, M., Litsyn, S., and Ron, D., Testing low-degree polynomials over GF(2). In *Approximation, randomization, and combinatorial optimization*, Lecture Notes in Comput. Sci. 2764, Springer-Verlag, Berlin 2003, 188–199.

[4] Alon, N., Kaufman, T., Krivelevich, M., Ron, D., Testing triangle-freeness in general graphs. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM Press, New York 2006, 279–288.

[5] Alon, N., and Shapira, A., A characterization of the (natural) graph properties testable with one-sided error. In *Proceedings of the 46th Annual Symposium on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2005, 429–438.

[6] Arora, S., Lund, C., Motwani, R., Sudan, M., and Szegedy, M., Proof verification and the hardness of approximation problems. *J. ACM* 45 (3) (1998), 501–555.

[7] Arora, S., and Sudan, M., Improved low degree testing and its applications. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 1997, 485–495.

[8] Batu, T., Dasgupta, S., Kumar, R., and Rubinfeld, R., The complexity of approximating the entropy. In *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 2002, 678–687.

[9] Batu, T., Fischer, E., Fortnow, L., Kumar, R., Rubinfeld, R., and White, P., Testing random variables for independence and identity. In *Proceedings of the 42nd Conference on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2001, 442–451.

[10] Batu, T., Fortnow, L., Rubinfeld, R., Smith, W., and White, P., Testing that distributions are close. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2000, 259–269.

[11] Batu, T., Kumar, R., and Rubinfeld, R., Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 2004, 381–390.

[12] Bellare, M., Coppersmith, D., Håstad, J., Kiwi, M., and M. Sudan, M., Linearity testing over characteristic two. *IEEE Trans. Inform. Theory* **42** (6) (1996), 1781–1795.

[13] Ben-Or, M., Coppersmith, D., Luby, M., and Rubinfeld, R., Non-abelian homomorphism testing, and distributions close to their self-convolutions. In *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*, Lecture Notes in Comput. Sci. 3122, Springer-Verlag, Berlin 2004, 273–285.

[14] Blum, M., Luby, M., and Rubinfeld, R., Self-testing/correcting with applications to numerical problems. *J. Comput. System Sci.* **47** (1993), 549–595.

[15] Bogdanov, A., Obata, K., and Trevisan, L., A lower bound for testing 3-colorability in bounded-degree graphs. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, IEEE Comput. Soc. Press, Los Alamitos, CA, 2002, 93–102.

[16] Coppersmith, D., Manuscript, 1989.

[17] Ergün, F., Kannan, S., Kumar, S. R., Rubinfeld, R., and Viswanathan, M., Spot-checkers. *J. Comput. System Sci.* **60** (3) (2000), 717–751.

[18] Ergün, F., Kumar, R., and Rubinfeld, R., Checking approximate computations of polynomials and functional equations. *SIAM J. Comput* **31** (2) (2001), 550–576.

[19] Fischer, E., The art of uninformed decisions: A primer to property testing. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **75** (2001), 97–126.

[20] Fischer, E., On the strength of comparisons in property testing. In *Electronic Colloquium on Computational Complexity* 8 (20), 2001.

[21] Fischer, E., and Matsliah, A., Testing graph isomorphism. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM Press, New York 2006, 299–308.

[22] Goldreich, O., Combinatorial property testing - a survey. In *Randomization Methods in Algorithm Design*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 43, Amer. Math. Soc., Providence, RI, 1999, 45–60.

[23] Goldreich, O., Goldwasser, S., and Ron, D., Property testing and its connection to learning and approximation. *J. ACM* **45** (4) (1998), 653–750.

[24] Goldreich, O., and Ron, D., On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity* 7 (20), 2000.

[25] Goldreich, O., and Ron, D., Property testing in bounded degree graphs. *Algorithmica* **32** (2002), 302–343.

[26] Goldreich, O., and Trevisan, L., Three theorems regarding testing graph properties. In *Proceedings of the 42nd Conference on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2001, 460–469.

[27] Jutla, C. S., Patthak, A. C., Rudra, A., and Zuckerman, D., Testing low-degree polynomials over prime fields. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2004, 423–432.

[28] Kaufman, T., Krivelevich, M., Ron, D., Tight bounds for testing bipartiteness in general graphs. *SIAM J. Comput.* **33** (2004), 1441–1483.

[29] Kaufman, T., and Ron, D., Testing polynomials over general fields. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2004, 413–422.

[30] Kiwi, M., Magniez, F., and Santha, M., Approximate testing with error relative to input size. *J. Comput. System Sci.* **66** (2) (2003), 371–392.

[31] Parnas, M., and Ron, D., Testing the diameter of graphs. *Random Structures Algorithms* **20** (2) (2002), 165–183.

[32] Polischuk, A., and Spielman, D., Nearly linear-size holographic proofs. In *Proceedings of the 26th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 1994, 194–203.

[33] Raskhodnikova, S., Ron, D., Rubinfeld, R., Shpilka, A., and Smith, A., Sublinear algorithms for string compressibility and the distribution support size. *Electronic Colloquium on Computational Complexity* 5 (125), 2005.

[34] Raz, R., and Safra, S., A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York 1997, 475–484.

[35] Ron, D., Property testing. In *Handbook of randomized computing*, Vol. II, Comb. Optim., Kluwer Acad. Publ., Dordrecht 2001, 597–649.

[36] Rubinfeld, R., On the robustness of functional equations. *SIAM J. Comput* **28** (6) (1999), 1972–1997.

[37] Rubinfeld, R., and Sudan, M., Robust characterization of polynomials with applications to program testing. *SIAM J. Comput.* **25** (2) (1996), 252–271.

Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, U.S.A.

E-mail: ronitt@csail.mit.edu

# Pseudorandomness and combinatorial constructions

Luca Trevisan*

**Abstract.** In combinatorics, the probabilistic method is a very powerful tool to prove the existence of combinatorial objects with interesting and useful properties. Explicit constructions of objects with such properties are often very difficult, or unknown. In computer science, probabilistic algorithms are sometimes simpler and more efficient than the best known deterministic algorithms for the same problem.

Despite this evidence for the power of random choices, the computational theory of pseudorandomness shows that, under certain complexity-theoretic assumptions, every probabilistic algorithm has an efficient deterministic simulation and a large class of applications of the probabilistic method can be converted into explicit constructions.

In this survey paper we describe connections between the conditional "derandomization" results of the computational theory of pseudorandomness and unconditional explicit constructions of certain combinatorial objects such as error-correcting codes and "randomness extractors."

**Mathematics Subject Classification (2000).** Primary 68Q10; Secondary 05D40.

**Keywords.** Computational complexity, pseudorandomness, derandomization, randomness extraction.

## 1. Introduction

**1.1. The probabilistic method in combinatorics.** In extremal combinatorics, the *probabilistic method* is the following approach to proving existence of objects with certain properties: prove that a random object has the property with positive probability. This simple idea has been amazingly successful, and it gives the best known bounds for most problems in extremal combinatorics. The idea was introduced (and, later, greatly developed) by Paul Erdős [18], who originally applied it to the following question: define $R(k, k)$ to be the minimum value $n$ such that every graph on $n$ vertices has either an independent set of size at least $k$ or a clique of size at least $k$.[1] It was known that $R(k, k)$ is finite and that it is at most $4^k$, and the question was to prove a lower bound. Erdős proved that a random graph with $2^{k/2}$ vertices has a positive probability of having no clique and no independent set larger than $k$, and so

*The author is supported in part by NSF grant CCF 0515231.

[1] Here by "graph" we mean an undirected graph, that is, a pair $G = (V, E)$ where $V$ is a finite set of *vertices* and $E$ is a set of pairs of elements of $E$, called *edges*. A clique in a graph $G = (V, E)$ is a set $C \subseteq V$ of vertices such that $\{u, v\} \in E$ for every two vertices $u, v \in C$. An independent set is a set $I \subseteq V$ of vertices such that $\{u, v\} \notin E$ for every two vertices $u, v \in I$.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$R(k,k) \geq 2^{k/2}$. The method, of course, gives no indication of how to actually construct a large graph with no small clique and no small independent set. Remarkably, in the past 60 years, there has been no asymptotic improvement to Erdős's lower bound and, perhaps more significantly, the best *explicit construction* of a graph without a clique of size $k$ and without an independent set of size $k$ has only about $k^{\log k}$ vertices [20], a bound that has not been improved in 25 years.

Shannon [59] independently applied the same idea to prove the existence of encoding schemes that can optimally correct from errors in a noisy channel and optimally compress data. The entire field of information theory arose from the challenge of turning Shannon's non-constructive results into algorithmic encoding and decoding schemes. We will return to the problem of encodings for noisy channels in Section 2. Around the same time, Shannon [60] applied the probabilistic method to prove the existence of boolean functions of exponential circuit complexity (see Section 4). Proving that certain specific boolean functions (for example, satisfiability of boolean formulae or 3-colorability of graphs) require exponential size circuits is a fundamental open problem in computational complexity theory, and little progress has been made so far.

The probabilistic method has found countless applications in the past 60 years, and many of them are surveyed in the famous book by Alon and Spencer [4].

For most problems in extremal combinatorics, as well as in information theory and complexity theory, probabilistic methods give the best known bound, and explicit constructions either give much worse bounds, or they give comparable bounds but at the cost of technical tours de force.

**1.2. Probabilistic methods in computer science.** In computer science, an important discovery of the late 1970s was the power of *probabilistic algorithms*. The most famous (and useful) of such algorithms are probably the polynomial time probabilistic algorithms for testing primality of integers [65], [47]. In these algorithms one looks for a "certificate" that a given number $n$ is composite; such a certificate could be for example an integer $a$ such that $a^n \not\equiv a \pmod{n}$, or four distinct square roots $\pmod{n}$ of the same integer. Rabin, Solovay and Strassen [65], [47] proved that there is a good chance of finding such certificates just by picking them at random, even though no efficient method to deterministically construct them was known.[2] Two other important and influential algorithms were discovered around the same time: an algorithm to test if two implicitly represented multivariate polynomials are identical [85], [56] (evaluate them at a random point chosen from a domain larger than the degree) and an algorithm to check if two vertices in a graph are connected by a path [2] (start a random walk at the first vertex, and see if the second vertex is reached after a bounded number of steps).[3]

[2] Note the similarity with the probabilistic method.

[3] The algorithm of Aleliunas et al. [2] broke new grounds in terms of *memory* use, not running time. It was already known that the Depth-First-Search algorithm could be used to solve the problem using linear time and a *linear* amount of memory. The random walk algorithm, however, needs only $O(\log |V|)$ bits of memory, and

A different type of probabilistic algorithms was developed starting in the late 1980s with the work of Sinclair and Jerrum [61]. These algorithms solve approximate "counting" problems, where one wants to know the number of solutions that satisfy a given set of combinatorial constraints. For example, given a bipartite graph, one would like to know, at least approximately, how many perfect matchings there are.[4] Sinclair and Jerrum introduced an approach based on a reduction to the problem of approximately sampling from the uniform distribution of all possible solutions. Since the latter problem involves randomness in its very definition, this approach inevitably leads to probabilistic algorithms.

**1.3. The computational theory of pseudorandomness.** In light of such algorithmic results, it was initially conjectured that probabilistic algorithms are strictly more powerful than deterministic ones and that, for example, there exist problems that can be solved probabilistically in polynomial time but that cannot be solved in polynomial time using deterministic algorithms.

This belief has been overturned by developments in the computational theory of pseudorandomness. The theory was initiated by Blum [9], Goldwasser and Micali [22], and Yao [84], with the motivation of providing sound foundations for cryptography. From the very beginning, Yao [84] realized that the theory also provides conditional *derandomization* results, that is, theorems of the form

> "if assumption X is true, then every problem that can be solved by a probabilistic polynomial time algorithm can also be solved by a deterministic algorithm of running time Y."

Yao showed that we can take X to be "there is no polynomial time algorithm that on input a random integer finds its prime factorization"[5] and Y to be "time $2^{n^\varepsilon}$ for every $\varepsilon > 0$."

An important project in complexity theory in the 1980s and 1990s was to strengthen Y to be "polynomial time" with a plausible X. The goal was achieved 1997 in a landmark paper by Impagliazzo and Wigderson [31], building on a considerable body of previous work.

At a very high level, the Impagliazzo–Wigderson result is proved in two steps. The first step (which is the new contribution of [31]) is the proof that an assumption about the worst-case complexity of certain problems implies a seemingly stronger assumption about the average-case complexity of those problems. A result of this kind is called an *amplification of hardness* result, because it "amplifies" a worst-case hardness assumption to an average-case one. The second step, already established

exponential improvement.

[4] A bipartite graph is a triple $G = (U, V, E)$ where $U$, $V$ are disjoint sets of vertices and $E \subseteq U \times V$ is a set of edges. A perfect matching is a subset $M \subseteq E$ such that for every $u \in U$ there is precisely one $v \in V$ such that $(u, v) \in M$, and vice versa.

[5] More generally, Yao showed that X can be "one-way permutations exist," see 5 for more details. The assumption about integer factorization implies the existence of one-way permutations, provided that we restrict ourselves to "Blum integers."

ten years earlier by Nisan and Wigderson [44], is the proof that the average-case assumption suffices to construct a certain very strong pseudorandom generator, and that the pseudorandom generator suffices to simulate deterministically in polynomial time every polynomial time probabilistic algorithm.

In conclusion, assuming the truth of a plausible complexity-theoretic assumption, every polynomial time probabilistic algorithm can be "derandomized," including the approximation algorithms based on the method of Sinclair and Jerrum. Furthermore, under the same assumption, a large class of applications of the probabilistic method in combinatorics can be turned into explicit constructions. We give some more details about the Impagliazzo–Wigderson Theorem in Section 6. The reader is also referred to the excellent survey paper of Impagliazzo [28] in the proceedings of the last ICM.

It is remarkable that many of the "predictions" coming from this theory have been recently validated unconditionally: Agrawal et al. [1] have developed a deterministic polynomial time algorithm for testing primality and Reingold [53] has developed a deterministic $O(\log n)$ memory algorithm for undirected graph connectivity. One can read about such developments elsewhere in these proceedings.

Here is an example of a question that is still open and that has a positive answer under the complexity-theoretic assumption used in the Impagliazzo–Wigderson work:

- Is there a deterministic algorithm that, on input an integer $n$, runs in time polynomial in $\log n$ and return a prime between $n$ and $2n$?

**1.4. When randomness is necessary.** Suppose that, in a distant future, someone proves the assumption used in the Impagliazzo–Wigderson work, so that we finally have an unconditional polynomial time derandomization of all probabilistic algorithms. Would this be the end of the use of randomness in computer science? The answer is no, for at least two reasons.

One reason is that such derandomization would probably not be practical. At a broad qualitative level, we consider polynomial-time algorithms as "efficient" and super-polynomial-time algorithms as "inefficient," and then such a result would establish the deep fact that "efficient" probabilistic algorithms and "efficient" deterministic algorithms have the same power. If the derandomization, however, causes a considerable (albeit polynomial) slow-down, and if it turns a practical probabilistic algorithm into an impractical deterministic one, then the probabilistic algorithm will remain the best choice in applications.

A more fundamental reason is that there are several applications in computer science where the use of randomness is *unavoidable*. For example, consider the task of designing a secure cryptographic protocol in a setting where all parties behave deterministically.

These observations lead us to consider the problem of generating randomness to be used in probabilistic algorithms, cryptographic protocols, and so on. Such generation begins by measuring a physical phenomenon that is assumed to be unpredictable (such as a sequence of physical coin flips) and that will be called a *random source* in

the following. Typically, one has access to random sources of very poor quality, and converting such measurements into a sequence of independent and unbiased random bits is a difficult problem. In Section 8 we discuss various approaches and impossibility results about this problem, leading to the definition of *seeded randomness extractor* due to Nisan and Zuckerman [45], [86].

Seeded randomness extractors have an amazing number of applications in computer science, often completely unrelated to the original motivation of extracting random bits from physical sources. They are related to hash functions, to pseudorandom graphs, to error-correcting codes and they are useful in complexity theory to prove, among other things, negative results for the approximability of optimization problems.

Ironically, the problem of generating high-quality random bits for cryptographic application is not satisfactorily solved by seeded randomness extractors (even though it was the original motivation for the research program that led to their definition). *Seedless* randomness extractors are needed for such application, and their theory is still being developed.

**1.5. Connections.** So far, we have discussed (i) the power of probabilistic methods, (ii) the *conditional* results proving that all probabilistic algorithms have a polynomial-time derandomization under complexity assumption, and (iii) the use of seeded randomness extractors to *unconditionally* run probabilistic algorithms in a setting in which only a weak source of randomness is available.

In Section 8 we describe a recently discovered *connection* between (ii) and (iii) and, more generally, between *conditional* results proved in the computational theory of pseudorandomness and *unconditional* explicit constructions of combinatorial objects.

One connection is between error-correcting codes and "hardness amplification" results. This connection has led to the application of coding-theoretic techniques in the study of average-case complexity. It is also possible to use complexity-theoretic techniques to build error-correcting codes, but so far this approach has not been competitive with previously known coding-theoretic techniques.

The second connection is between pseudorandom generators and seeded randomness extractors. This connection has led to improvements in both settings.

Various impossibility results are known for error-correcting codes and randomness extractors. Via these connections, they imply impossibility results for hardness amplification and conditional derandomization. In Section 7 we discuss approaches to sidestep these negative results.

## 2. Pseudorandom objects: codes and graphs

In this section we introduce two examples of very useful combinatorial objects whose existence easily follows from the probabilistic method: error-correcting codes and expander graphs. Explicit constructions of such objects are also known.

**2.1. Error-correcting codes.** Consider the process of picking a random set $S \subseteq \{0,1\}^n$ of size $2^k$, $k < n$. If, say, $k = n/2$, then it is easy to show that there is an absolute constant $\delta > 0$ such that, with high probability, every two elements $u, v \in S$ differ in at least $\delta n$ coordinates. By a more careful estimate, we can also see that there is an absolute constant $c$ such that, for every $\varepsilon > 0$, it is likely that every two elements of $S$ differ in at least $(1/2 - \varepsilon)n$ coordinates with high probability, provided $k \leq c\varepsilon^2 n$. For reasons that will be clear shortly, let us change our perspective slightly, and consider the (equivalent) process of picking a random injective function $C\colon \{0,1\}^k \to \{0,1\}^n$: clearly the same bounds apply.

For two strings $u, v \in \{0,1\}^n$, the Hamming distance between $u$ and $v$ (denoted $d_H(u,v)$) is the number of coordinates where $u$ and $v$ differ, that is

$$d_H(u,v) := |\{i : u_i \neq v_i\}|. \tag{1}$$

**Definition 2.1** (Error-correcting code). We say that $C\colon \{0,1\}^k \to \{0,1\}^n$ is an $(n,k,d)$-*code* if $d_H(C(x), C(y)) \geq d$ for every two distinct $x, y \in C$.

This concept is due to Hamming [25]. Error-correcting codes are motivated by the following scenario. Suppose we, the *sender*, have a $k$-bit message $M \in \{0,1\}^k$ that we want to transmit to a *receiver* using an unreliable *channel* that introduces errors, and suppose we have an $(n,k,d)$-code $C$. Then we can compute $c = C(M)$ and transmit $c$ over the channel. The receiver gets a string $c'$, which is a corrupted version of $c$, and looks for the message $M'$ that minimizes $d_H(C(M'), c')$. If the channel introduces fewer than $d/2$ errors, then the receiver correctly reconstructs $M$.[6]

Keeping this application in mind, for every given $k$, we would like to construct $(n,k,d)$-codes where $d$ is as large as possible (because then the receiver can tolerate more errors) and $n$ is as small as possible (so that we do not have to communicate a very long encoding). Furthermore, we would like $C$ and the decoding procedure run by the receiver to be computable by efficient algorithms.

One trade-off between the parameters is that $\frac{d}{n} \leq \frac{1}{2} + o_k(1)$. Keeping in mind that the number of errors that can be corrected is at most $d/2$, this means that the receiver can correctly reconstruct the message only if the number of errors is at most $\left(\frac{1}{4} + o_k(1)\right) n$.

It is possible to do better if we are willing to settle for the notion of "list-decodability," introduced by Elias [17].

**Definition 2.2** (List-decodable code). We say that $C\colon \{0,1\}^k \to \{0,1\}^n$ is $(L,\delta)$-*list decodable* if for every $u \in \{0,1\}^n$,

$$|\{x \in \{0,1\}^k : d_H(C(x), u) \leq \delta n\}| \leq L.$$

[6] It is also easy to see that this analysis is tight. If there are two messages $M$, $M'$ such that $d_H(C(M), C(M')) = d$ and we send $M$, then it is possible that even a channel that introduces only $d/2$ errors can fool the receiver into thinking that we sent $M'$.

Here the idea is that we send, as before, the encoding $C(M)$ of a message $M$. The receiver gets a string $u$ and computes the *list* of all possible messages $M'$ such that $d_H(C(x), u) \leq \delta n$. If $C$ is an $(L, \delta)$-code, then the list is guaranteed to be of length at most $L$, and if the channel introduces at most $\delta n$ errors then our message is guaranteed to be in the list.

Using the probabilistic method, it is easy to show the existence of $(L, 1/2-\varepsilon)$-list decodable codes $C\colon \{0,1\}^k \to \{0,1\}^n$ for every $k$ and $\varepsilon$, where $n = O(k\varepsilon^{-2})$ and $L = O(\varepsilon^{-2})$. It was also known how to define *efficiently encodable* codes with good (but not optimal) parameters. It took, however, 40 years until Sudan [66] defined the first *efficient list-decoding* algorithm for such codes. Sudan's algorithm suffices to define $(\varepsilon^{-O(1)}, 1/2-\varepsilon)$-list decodable codes $C\colon \{0,1\}^k \to \{0,1\}^n$ with $n = (k/\varepsilon)^{O(1)}$ for every $k$, $\varepsilon$, and the codes are encodable and list-decodable in time polynomial in $n$. This means that even if the channel introduces close to $n/2$ errors, it is still possible for the receiver to gain considerable information about the message. (Namely, the fact that the message is one out of a small list of possibilities.) Other list-decoding algorithms are now known, but they are beyond the scope of this survey. See Sudan's survey [67], Guruswami's thesis [23] and two recent breakthrough papers [24], [46].

**2.2. Expander graphs.** Consider the process of picking at random a graph according to the $G_{n,\frac{1}{2}}$ distribution.(The $G_{n,\frac{1}{2}}$ distribution is the uniform distribution over the set of $2^{\binom{n}{2}}$ graphs over $n$ vertices.) A simple calculation shows that for every two disjoint sets of vertices $A$, $B$ there are $\left(\frac{1}{2} \pm o_n(1)\right)|A||B|$ edges with one endpoint in $A$ and one endpoint in $B$. Chung, Graham and Wilson [15] call a family of graphs satisfying the above properties a family of *quasi-random* graphs, and prove that six alternative definitions of quasi-randomness are all equivalent. Explicit constructions of quasi-random graphs are known, and the notion has several applications in combinatorics. (See the recent survey paper by Krivelevich and Sudakov [37].) Consider now a process where we randomly generate an $n$-vertex graph where every vertex has degree at most $d$ (think of $d$ as a fixed constant and $n$ as a parameter). For example, consider the process of picking $d$ perfect matchings and then taking their union. Then it is possible to show that for every two disjoint sets of vertices $A$, $B$ there are $(1 \pm o_{n,d}(1))d\frac{|A||B|}{n}$ edges with one endpoint in $A$ and one endpoint in $B$. (Families of) graphs with this property are called *expanders*, and they have several applications in computer science. To gain a sense of their usefulness, imagine that an expander models a communication network and note that if $o(dn)$ edges are deleted, the graph still has a connected component with $(1-o(1))n$ vertices. Furthermore, expander graphs have several other interesting properties: they have small diameter, it is possible to find several short edge-disjoint paths between any two vertices, and so on. There are other possible definitions of expanders, which are related but not equivalent. In one possible (and very useful) definition, expansion is measured in terms of the *eigenvalue gap* of the *adjacency matrix* of the graph (see e.g. the discussion in [37]). For this definition, Lubotzky, Phillips and Sarnak [41] provide an optimal explicit construction. Another possible measure is the *edge expansion* of the graph. Optimal

explicit constructions for this measure are not known, but considerable progress is made in [13], [3].

## 3. Randomness extractor

Randomness extractors are procedures originally designed to solve the problem of generating truly random bits. As we will see, randomness extractors can be seen as a sort of pseudorandom graphs, they can be constructed using techniques from the field of pseudorandomness, and they are tightly related to constructions error-correcting codes, expanders and other random-like combinatorial objects.

**3.1. Generating random bits.** In order to generate random bits in practice, one starts by measuring a physical phenomenon that is assumed to contain randomness.[7] For example, in many computer systems one starts by collecting statistics on the user's keystrokes or mouse movement, or on the latency time of disk access, and so on. This raw data, which is assumed to contain some amount of entropy, is then passed to a "hash function," and the output of the function is assumed to be a sequence of truly random bits. Such systems, widely used in practice, are typically not validated by any rigorous analysis.

In a mathematical modeling of this situation, we have a random variable $X$ representing our physical measurement, ranging, say, over $\{0,1\}^n$. We would like to construct a function $\mathrm{Ext}\colon \{0,1\}^n \to \{0,1\}^m$ such that, by making as little assumptions on $X$ as possible, we can prove that $\mathrm{Ext}(X)$ is distributed uniformly over $\{0,1\}^m$, or at least it is approximately so.

Von Neumann [82] studied a version of this problem where $X$ is a sequence of independent and identically distributed biased coin tosses. The independence assumption is crucially used. The general problem was extensively studied in computer science in the 1980s [55], [80], [79], [78], [14], [16]. Notably, the goal was to define a single function Ext that would work for as large as possible a class of distributions $X$. An early conclusion was that the extraction problem is impossible [55], as defined above, even if just very weak forms of dependencies between different bits are allowed in the distribution of $X$. Two approaches have been considered to circumvent this impossibility.

1. One approach is to consider a model with a *small* number of mutually independent random variables $X_1, \ldots, X_k$, each satisfying weak randomness requirements. This line of work, initiated in [55], [80], [14], saw no progress for a long time, until recent work by Barak et al. [7] made such progress possible by a breakthrough in additive combinatorics [12], [36]. This is now a very active area of research [8], [53], [11], [87], [48] with connections to other areas of combinatorics.

[7] We will not get into the physical and philosophical problems raised by such assumption.

2. The other approach, initiated in [79], is to stick to the model of a single sample $X$ and to consider the following question: suppose we have a randomized algorithm $A$ (that is correct with high probability given the ability to make truly random choices) and suppose we have an input $x$. Can we efficiently find what is the most probable output of $A(x)$?

**3.2. The definition of randomness extractors.** To formalize approach (2) it is convenient to think of a probabilistic algorithm $A(\cdot, \cdot)$ as having two inputs: a "random" input $r$ and a "regular" input $I$. We say that "$A$ computes a function $f$ with high probability" if, for every $I$,

$$\mathbb{P}[A(r, I) = f(I)] \geq .9$$

where the probability is taken with respect to the uniform distribution over bit strings $r$ of the proper length.[8]

Let $U_n$ denote a random variable uniformly distributed over $\{0, 1\}^n$.

Suppose that our algorithm $A$ requires $m$ truly random bits to process a given input $x$. Furthermore, suppose that we can define a function $\mathrm{Ext}\colon \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^m$ such that if $X$ is our physical source and $U_d$ is uniformly distributed over $\{0, 1\}^d$ then $\mathrm{Ext}(X, U_d)$ is uniformly distributed over $\{0, 1\}^m$. Here is a way to simulate $A(\cdot)$ using $X$: (i) get a sample $x \sim X$, (ii) for every $s \in \{0, 1\}^d$, compute $a_s := A(\mathrm{Ext}(x, s), I)$, (iii) output the most common value among the $a_s$.

It is now easy to show that the above algorithm computes $f(I)$ with probability at least .8, over the choice of $X$. This is because $\mathbb{P}[A(\mathrm{Ext}(U_d, X), I) = f(I)] \geq .9$ and so

$$\mathop{\mathbb{P}}_{X}\Big[\mathop{\mathbb{P}}_{U_d}[A(\mathrm{Ext}(U_d, X), I) = f(I)] > \tfrac{1}{2}\Big] \geq .8. \tag{2}$$

The running time of our simulation of $A$ is $2^d$ times the running time of $A$, which is polynomial in the running time of $A$ provided that $d$ is logarithmic.

For this reasoning to work it is not necessary that $\mathrm{Ext}(X, U_d)$ be distributed *exactly* uniformly, but it is enough if it approximates the uniform distribution in an appropriate technical sense. If $X$ and $Y$ are two random variables taking values in $\Omega$, then we define their *variational distance* (also called *statistical distance*) as

$$\|X - Y\|_{\mathrm{SD}} := \max_{T \subseteq \Omega} |\,\mathbb{P}[X \in T] - \mathbb{P}[Y \in T]|. \tag{3}$$

We will sometimes call sets $T \subseteq \Omega$ *statistical tests*. If $\|X - Y\|_{\mathrm{SD}} \leq \varepsilon$ then we say that $X$ is $\varepsilon$-close to $Y$.

We say that $\mathrm{Ext}\colon \{0, 1\}^n \to \{0, 1\}^d \to \{0, 1\}^m$ is a *seeded extractor* for a distribution $X$ with *error parameter* $\varepsilon$ if $\mathrm{Ext}(X, U_d)$ is $\varepsilon$-close to $U_m$.

[8] The reader may find .9 to be a poor formalization of the notion of "with *high* probability," but it is easy to reduce the error probability at the cost of a moderate increase of the running time.

Vazirani and Vazirani [79] provided extractors for a certain class of distributions. (Their terminology was different.) Zuckerman [86] was the first to show that extractors exist for a very general class of distributions. Define the min-entropy of $X$ as $H_\infty(X) := \min_a \log_2 \frac{1}{\mathbb{P}[X=a]}$. If $H_\infty(X) \geq k$, then we say that $X$ is a $k$-source.

**Definition 3.1.** A function $\mathrm{Ext}\colon \{0,1\}^n \to \{0,1\}^d \to \{0,1\}^m$ is a $(k,\varepsilon)$ *seeded extractor* if $\mathrm{Ext}(X, U_d)$ is $\varepsilon$-close to $U_m$ for every $k$-source $X$.

The definition is implicit in [86]. The term extractor was coined in [45]. The term "seeded" refer to the truly random input of length $d$, which is called a *seed*. From now, we will refer to seeded extractors as simply "extractors."

Let $\mathrm{Ext}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a $(k,\varepsilon)$-extractor. Construct a bipartite graph $G = ([N],[M],E)$ with $N = 2^n$ vertices on the left, $M = 2^m$ vertices on the right. Connect two vertices $u, v$ if there is an $s$ that $v = \mathrm{Ext}(u,s)$. Then if we pick any subset $S \subseteq [N]$ on the left and any subset $T \subseteq [M]$ on the right, the number of edges is $|S| \cdot 2^d \cdot |T|/2^m$ plus or minus $\varepsilon|S|2^d$, provided $|S| \geq 2^k$. This is similar to one of the definitions of expander. Zuckerman and Wigderson [83] prove that one can derive expanders with very strong "edge expansion" from extractors.

Radakrishnan and Ta-Shma show that, in every extractor, $d \geq \log(n-k) + 2\log(1/\varepsilon) - O(1)$ and that $m \leq k + d - O(1)$. Non-constructively, one can show that such bounds are achievable up to the additive constant factor, but explicit constructions are difficult. We will discuss explicit constructions later.

**3.3. Applications.** Randomness extractors have several applications, some of which are described below. See the tutorial by Salil Vadhan [76] and the survey by Ronen Shaltiel [57] for more examples and a broader discussion.

**Simulation of randomized algorithms.** Suppose we have a randomized algorithm $A$ that on input $I$ computes $f(I)$ with probability, say, .9, and suppose that Ext is a $(k', 1/4)$-extractor and that $X$ is a $k$-source. As before, let us sample $x \sim X$ and compute $A(\mathrm{Ext}(x,s), I)$ for every $s$ and output the majority value. Let $B$ be the set of $x$ such that the algorithm fails. If $|B| \geq 2^{k'}$, then consider a random variable $Y$ uniformly distributed over $B$. It has entropy $k$, so $\mathrm{Ext}(Y, U_d)$ should be $1/4$-close to uniform. Consider the statistical test $T$ defined as

$$T := \{r : A(r,I)\} = f(I)\}. \tag{4}$$

Then $\mathbb{P}[U_n \in T] \geq .9$ by assumption and $\mathbb{P}[\mathrm{Ext}(Y, U_d) \in T] \leq 1/2$ by construction. This would contradict Ext being an extractor. We then conclude that $|B| \leq 2^{k'}$, and so the probability that our algorithm fails is at most $\mathbb{P}[X \in B] \leq |B|/2^k \leq 2^{k'-k}$.

This is very useful even in a setting in which we assume access to a perfect random source. In such a case, by using $n$ truly random bits we achieve an error probability that is only $2^{k'-n}$. Note that, in order to achieve the same error probability by running the algorithm several times independently we would have used $O((n-k')\cdot m)$ random bits instead of $n$.

**Other applications.** Randomness extractors are also very useful in settings where we assume a fully random distribution, say, over $n$ bits, that is unknown to us, except for some partial information of entropy at most $n - k$ bits.

Then the distribution of the unknown string *conditioned on our knowledge* still has entropy at least $k$. If an extractor is applied to the unknown string, then the output of the extractor will be uniformly distributed even conditioned on our knowledge. In other words, our knowledge in useless in gaining any information about the output of the extractor.

This approach is used in the cryptographic settings of *privacy amplification* and *everlasting security* and in the design of pseudorandom generators for space-bounded algorithms. See [39], [77] and the references therein for the application to everlasting security and [45], [30], [51] for the application to pseudorandom generators.

## 4. Circuit complexity

In order to discuss the computational approach to pseudorandomness we need to define a measure of efficiency for algorithms. We will informally talk about the "running time" of an algorithm on a given input without giving a specific definition. The reader can think of it as the number of elementary operations performed by an implementation of the algorithm on a computer. A more formal definition would be the number of steps in a Turing machine implementation of the algorithm. (See e.g. [64] for a definition of Turing machine.)

We say that a set $L \subseteq \{0,1\}^*$ is *decidable* in time $t(n)$ if there is an algorithm that on input $x \in \{0,1\}^n$ decides in time $\leq t(n)$ whether $x \in L$.

We are also interested in a more "concrete" measure of complexity, called *circuit complexity*. For integers $n$ and $i \leq n$, define the set $P_{i,n} := \{(a_1, \ldots, a_n) \in \{0,1\}^n : a_i = 1\}$. We say that a set $S \subseteq \{0,1\}^n$ has a *circuit of size* $K$ if there is a sequence of sets $S_1, \ldots, S_K$ such that: (i) $S_K = S$ and (ii) each $S_j$ is either a set $P_{i,n}$, or it is the complement of a set $S_h$, $h < j$, or it is the union $S_h \cup S_\ell$ of two sets, with $h, \ell < j$ or it is the intersection $S_h \cap S_\ell$ of two sets, with $h, k < j$. We say that a function $f : \{0,1\}^n \to \{0,1\}$ has a circuit of size $K$ if it is the characteristic function of a set that has a circuit of size $K$.

The *circuit complexity* of a set $S$ is the minimum $K$ such that $S$ has a circuit of size $K$. (Similarly for boolean functions.)

It is easy to see that there are subsets of $\{0,1\}^n$ whose circuit complexity is at least $c\frac{2^n}{n}$ for some constant $c > 0$: if a set has circuit complexity at most $K$, then it can be described by using only $O(K \log K)$ bits, and so there are $2^{O(K \log K)}$ sets of circuit complexity at most $K$. If this number is less than $2^{2^n}$ then there exists a set of circuit complexity larger than $K$. Indeed, by the same argument, a random set has circuit complexity at least $c\frac{2^n}{n}$ with very high probability.

If $L \subseteq \{0,1\}^*$ is a set decidable in time $t(n)$, then for every $n$ there is a circuit

of size $O((t(n))^2)$ for $L \cap \{0,1\}^n$. This implies that in order to prove lower bounds on the running time of algorithms for a given decision problem it is enough to prove lower bounds for the circuit complexity of finite fragments of it.[9]

So far there has been very little success in proving circuit complexity lower bounds for "explicit sets," such as sets in NP. The strongest known lower bound is $5n$ [38], [32], and even an $n \log n$ lower bound is considered hopelessly out of reach of current techniques.

This is perhaps surprising given the simplicity of the definition of circuit complexity. The definition looks like a finite version of the definition of complexity for Borel sets, and one may hope that one could transfer techniques from topology to this setting. Sipser describes this idea in [62], [63], but, unfortunately, so far it has not led to any lower bound for general circuits.

Complexity theorists' failure to prove strong circuit lower bounds is partly explained by a famous paper by Razborov and Rudich [52]. They describe a general class of approaches to lower bounds that they call "natural proofs." Razborov and Rudich show that all known methods to prove lower bounds for restricted classes of circuits yield natural proofs, but that (under certain complexity-theoretic assumptions) natural proofs cannot prove lower bounds for general circuits. The complexity theoretic assumption is itself about circuit lower bounds, and it is used to construct certain pseudorandom generators. The pseudorandom generators, in turn, imply the impossibility result. Somewhat inaccurately, the Razborov–Rudich result can be summarized as:

*Circuit lower bounds are difficult to prove because they are true.*

## 5. Pseudorandom generators and their application to derandomization

Informally, a pseudorandom generator is an efficiently computable map $G\colon \{0,1\}^t \to \{0,1\}^m$, where $m$ is much bigger than $t$, such that, for a uniformly selected $x \in \{0,1\}^t$, the distribution $G(x)$ is pseudorandom, that is, it "looks like" the uniform distribution over $\{0,1\}^m$. We begin by describing how to formalize the notion of a distribution "looking like" the uniform distribution, and, more generally, the notion of two distributions "looking like" one other.

Recall that we use $U_n$ to denote a random variable that is uniformly distributed in $\{0,1\}^n$.

Ideally, we would like to say that $G(\cdot)$ is a good pseudorandom generator if $G(U_t)$ and $U_m$ are close in statistical distance. Then, as we already discussed in Section 3, every application in which $m$ truly random bits are needed could be realized using the output of the generator (with a small increase in the probability of error). Unfortunately, this is too strong a definition: consider the statistical test $T$ defined to be the set of all possible outputs of $G$. Then $\mathbb{P}[G(U_t) \in T] = 1$ but $\mathbb{P}[U_m \in T] \leq 2^{t-m}$.

[9] The converse is not true: one can have undecidable sets of bounded circuit complexity.

The great idea that came from the work of Blum, Goldwasser, Micali and Yao in 1982 ([9], [22], [84]) was to modify the notion of statistical distance by considering only *efficiently computable* statistical tests.

**Definition 5.1** (Computational indistinguishability). Two distributions $\mu_X$ and $\mu_Y$ over $\{0,1\}^m$ are $(K,\varepsilon)$-*indistinguishable* if for every set $T \subseteq \{0,1\}^m$ of circuit complexity at most $K$,

$$\Big| \mathop{\mathbb{P}}_{x\sim\mu_X}[x \in T] - \mathop{\mathbb{P}}_{y\sim\mu_Y}[y \in T]\Big| \leq \varepsilon.$$

**Definition 5.2** (Pseudorandomness). A distribution $\mu_X$ over $\{0,1\}^m$ is $(K,\varepsilon)$-*pseudorandom* if it is $(K,\varepsilon)$-indistinguishable from the uniform distribution. That is, for every $T \subseteq \{0,1\}^m$ of circuit complexity $\leq K$,

$$\Big| \mathop{\mathbb{P}}_{x\sim\mu_X}[x \in T] - \tfrac{|T|}{2^m}\Big| \leq \varepsilon.$$

The following definition is due to Nisan and Wigderson [44].

**Definition 5.3** (Quick pseudorandom generator). Suppose that for every $n$ there is a $G_n\colon \{0,1\}^{t(n)} \to \{0,1\}^n$ that is $(n^2, 1/n)$-pseudorandom, and that there is an algorithm $G$ that, given $n$, $s$, computes $G_n(s)$ in time $2^{O(t(n))}$. Then $G$ is called a $t(n)$-*quick* pseudorandom generator.

Suppose that an $O(\log n)$-quick pseudorandom generator (abbreviated logQPRG) exists, and suppose that $f$ is a function and $A$ is a polynomial time randomized algorithm that computes $f$ with probability at least $3/4$. We now describe a derandomization of the algorithm $A$.

Let $I$ be an input, and let $m$ be the number of random bits used by $A$ on input $I$. Let $K$ be an efficiently computable upper bound for the circuit complexity pf $T := \{r : A(r,I) = f(I)\}$. Choose $n$ to be large enough so that: (i) $n^2 \geq K$, (ii) $n \geq m$, and (iii) $n \geq 5$. Because of our assumption that $A$ runs in polynomial time, $n$ is polynomial in the length of $I$.[10]

Now compute $A(G_n(s), I)$ for each $s$, and output the value that is returned most often. This completes the description of a polynomial time deterministic algorithm.

Regarding correctness, we assumed $\mathbb{P}[A(U_m, I) = f(I)] \geq \frac{3}{4}$, and so

$$\mathbb{P}[A(G_n(U_t(n)), I) = f(I)] \geq \tfrac{3}{4} - \tfrac{1}{n} > \tfrac{1}{2}. \tag{5}$$

Otherwise, the set $T = \{r : A(r,I) = f(I)\}$ contradicts the pseudorandomness of $G_n$. Something similar can be done if $A$ is only guaranteed to *approximate* $f$ with high probability, for example if $f(I)$ is the number of perfect matchings in the graph

[10] It should be noted that we may not know how to construct a circuit for $T$, because it seems that to construct such a circuit we need to know $f(I)$. In order to compute a polynomially bounded upper bound for the circuit complexity of $T$, however, we just need to find out how large the circuit for $T$ is that *we would be able to build if we knew* $f(I)$.

represented by $I$ and $A$ is the Jerrum–Sinclair–Vigoda probabilistic approximation algorithm for this problem [33]. The only difference is that we take the *median* of the outputs instead of the most common one.

The applications of logQPRGs to the probabilistic method is as follows. Suppose that:

- For every $n$, we have a set $\Omega_n$ of "objects of size $n$" (for example, graphs with $n$ vertices and maximum degree $d$, where $d$ is a fixed constant). It is convenient to assume the sets $\Omega_n$ to be disjoint.
- We define $P \subseteq \bigcup_n \Omega_n$ to be the set of *interesting* objects that we would like to construct. (For example, expander graphs.)
- Property $P$ is computable in polynomial time. That is, there is an algorithm that, given $n$ and $x \in \Omega_n$, runs in time polynomial in $n$ and determines whether $x \in P$.
- The probabilistic method proves that such graphs exist and are "abundant." That is, for every $n$, we define a probability distribution $\mu_n$ over $\Omega_n$ and we prove that $\mathbb{P}_{x\sim\Omega_n}[x \in P] \geq 1/2$. (The constant $1/2$ is not important.)
- The distributions $\mu_n$ are polynomial time samplable. That is, there is a probabilistic algorithm $A$ that, given $n$, generates in time polynomial in $n$ a sample from $\mu_n$.

This formalization captures the way the probabilistic method is typically used in practice, with the exception of the efficient computability of $P$, which sometimes is not true. (For example, in the problem of finding lower bounds for $R(k, k)$.) Finally, suppose that a logQPRG exists. Given $n$ here is how we construct an element in $P \cap \Omega_n$. Let $m$ be the number of random bits used by $A$ to sample an element of $\mu_n$, and let $K$ be an upper bound for the size of a circuit for the set $T := \{r : A(n, r) \in P\}$. As before, we can use the assumption that $A$ is computable in polynomial time and $P$ is decidable in polynomial time to conclude that $m$ and $K$ are upper bounded by polynomials in $n$. Let $N$ be large enough so that (i) $N \geq 3$, (ii) $N^2 \geq K$ and (iii) $N \geq m$. Then compute $A(n, G_N(N, s))$ for every $s$, and let $s_0$ be such that $A(n, G_N(N, s_0)) \in P$. Such an $s_0$ must exist, otherwise $T$ contradicts the pseudorandomness of $G_N$. Output $A(n, G_N(N, s_0))$.

## 6. Conditional constructions of pseudorandom generators

Blum and Micali [9] construct $n^{o(1)}$QPRGs, according to a slightly different definition, assuming a specific number-theoretic assumption. Yao [84] proves that the Blum–Micali definition is equivalent to a definition based on indistinguishability and

constructs $n^{o(1)}$QPRGs under the more general assumption that *one-way permutations* exist. Yao [84] also recognizes that $n^{o(1)}$QPRGs imply a $2^{n^{o(1)}}$ derandomization of every probabilistic algorithm.

Blum, Micali and Yao do not use the parametrization that we adopted in the definition of quick pseudorandom generators. In the cryptographic applications that motivate their work, it is important that the generator be computable in time polynomial in the length of the output (rather than exponential in the length of the input), and, if $m$ is the length of the output, one desires $(S(m), \varepsilon(m))$-pseudorandomness where $S(m)$ and $1/\varepsilon(m)$ are super-polynomial in $m$. Their constructions satisfy these stronger requirements.

Håstad et al. [26] show that the weaker assumption that *one-way functions* exist suffices to construct $n^{o(1)}$QPRGs. Their construction satisfies the stronger requirements of [9], [84]. We do not define one-way permutations and one-way functions here and we refer the interested reader to Goldreich's monograph [21], the definitive treatment of these results.

Nisan and Wigderson [44] introduced the definition of quick pseudorandom generator that we gave in the previous section and presented a new construction that works under considerably weaker assumptions than the existence of one-way functions.[11] The Nisan–Wigderson construction also "scales" very well, and it gives more efficient QPRGs if one is willing to start from stronger assumptions. A sufficiently strong assumption implies optimal logQPRGs, and this is the only version of the Nisan–Wigderson results that we will discuss.

We first need to define the notion of *average-case circuit complexity.* We say that a set $S \subseteq \{0,1\}^n$ is $(K, \varepsilon)$-hard on average if for every set $T$ computable by a circuit of size $\leq K$ we have $\mathbb{P}[1_S(x) = 1_T(x)] \leq \frac{1}{2} + \varepsilon$, where we use the notation $1_S$ for the characteristic function of the set $S$. We say that a set $L \subseteq \{0,1\}^*$ is $(K(n), \varepsilon(n))$-hard on average if, for every $n$, $L \cap \{0,1\}^n$ is $(K(n), \varepsilon(n))$-hard on average.

**Theorem 6.1** (Nisan and Wigderson [44]). *Suppose there is a set L such that:* (i) *L can be decided in time* $2^{O(n)}$ *and* (ii) *there is a constant* $\delta > 0$ *such that L is* $(2^{\delta n}, 2^{-\delta n})$*-hard on average. Then a* log*QPRG exists.*

When Theorem 6.1 was announced in 1988, average-case complexity was much less understood than worst-case complexity and it was not even clear if the assumption used in the theorem was plausible.

This motivated a long-term research program on average-case complexity. Building on work by Babai, Fortnow, Impagliazzo, Nisan and Wigderson [6], [27], Impagliazzo and Wigderson finally proved in 1997 that the assumption of Theorem 6.1 is equivalent to a seemingly weaker *worst-case* assumption.

[11] On the other hand, the Nisan–Wigderson generator does not satisfy the stronger properties of the pseudorandom generators of Blum, Micali, Yao, Håstad et al. [9], [84], [26]. This is unavoidable because the existence of such stronger pseudorandom generators is *equivalent* to the existence of one-way functions.

**Theorem 6.2** (Impagliazzo and Wigderson [31]). *Suppose there is a set $L$ such that:* (i) *$L$ can be decided in time $2^{O(n)}$ and* (ii) *there is a constant $\delta > 0$ such that the circuit complexity of $L$ is at least $2^{\delta n}$.*

*Then there is a set $L'$ such that:* (i) *$L'$ can be decided in time $2^{O(n)}$ and* (ii) *there is a constant $\delta' > 0$ such that $L'$ is $(2^{\delta' n}, 2^{-\delta' n})$-hard on average.*

In conclusion, we have optimal logQPRG and polynomial time derandomization of probabilistic algorithms under the assumptions that there are problems of exponential circuit complexity that are computable in exponential time. Such an assumption is considered very plausible.

There are other applications of these techniques that we will not have space to discuss, including extensions to the case of pseudorandomness against "non-deterministic statistical tests," which imply surprising results for the Graph Isomorphism problem [35], [43].

## 7. Average-case complexity and codes

We now come to a connection between the Impagliazzo–Wigderson Theorem and error-correcting codes. Due to space limitations we will only give a short discussion. The interested reader is referred to our survey paper [72] for more details.

Impagliazzo and Wigderson derive Theorem 6.2 from the following "hardness amplification" reduction.

**Theorem 7.1** (Impagliazzo and Wigderson [31]). *For every $\delta > 0$ there are constants $\delta' > 0$, $c > 1$, and an algorithm with the following property.*

*If $S \subseteq \{0,1\}^n$ is a set of circuit complexity at least $2^{\delta n}$, then, on input $S$, the algorithm outputs a set $S' \subseteq \{0,1\}^{cn}$ that is $(2^{\delta' n}, 2^{-\delta' n})$-hard on average.*

Like most results in complexity theory, the proof is by contradiction: suppose we have a set $T$ computable by a circuit of size $2^{\delta' n}$ such that $\mathbb{P}_{x \sim \{0,1\}^n}[1_{S'}(x) = 1_T(x)] \geq 1/2 + 2^{-\delta' n}$; then Impagliazzo and Wigderson show how to use such a circuit for $T$ to construct a circuit for $S$ of size $2^{\delta n}$.

Phrased this way, the result has a strong coding-theoretic flavor: we can think of $S$ as a "message," of $S'$ as the "encoding" of $S$, of $T$ as the "corrupted transmission" that the receiver gets, and of the process of reconstructing (a circuit for) $S$ from (a circuit for) $T$ as a "decoding" process. Given this perspective, introduced in [68], it is natural to try and apply coding-theoretic algorithms to hardness amplification. In doing so, we encounter the following difficulty: viewed as a message, a set $S \subseteq \{0,1\}^n$ is (or can be represented as) a bit-string of length $N = 2^n$, and so a polynomial time coding-theoretic algorithm that reconstructs $S$ from a corrupted encoding of $S$ takes time $N^{O(1)} = 2^{O(n)}$. In Theorem 7.1 however we need to produce a circuit of size $2^{\delta n} = N^{\delta}$, and so the circuit cannot simply be an implementation of the decoding algorithm.

It seems that what we need is the following type of error-correcting code (we use the notation $\mathcal{P}(A)$ to denote the set of all subsets of a set $A$): a map $C\colon \mathcal{P}(\{0,1\}^n) \to \mathcal{P}(\{0,1\}^{n'})$ with $n' = O(n)$ such that there is an algorithm that – given a set $T \in \mathcal{P}(\{0,1\}^{n'})$ close to the encoding $C(S)$ of a message $S \in \mathcal{P}(\{0,1\}^{n'})$ and an element $a \in \{0,1\}^n$ – determines in time at most $2^{\delta n}$ whether $a \in S$ or not. If we think of a set $S \in \mathcal{P}(\{0,1\}^n)$ as simply a bit-string in $\{0,1\}^N$, $N = 2^n$, then we are looking for an error correcting code $C\colon \{0,1\}^N \to \{0,1\}^{N'}$, $N' = N^{O(1)}$, such that there is an algorithm that, given a string $u \in \{0,1\}^{N'}$ close to an encoding $C(x)$ and given an index $i \in \{1,\ldots,N\}$, computes in time at most $N^\delta$ the bit $x_i$. It remains to specify how to "give in input" a string $u$ of length $N' > N$ to an algorithm of running time, say, $N^{.001}$: the algorithm does not even have enough time to *read* the input. This can be handled by modeling the input as an "oracle" for the algorithm, which is a standard notion.

The existence of error-correcting codes with this kind of "sub-linear time decoding algorithms" was well known, but the problem is that this notion is still not sufficient for the application to Theorem 7.1. The reason is that we have described a decoding algorithm that gives a unique answer and, as discussed in Section 2, such algorithms cannot recover from more than a $1/4 + o(1)$ fraction of errors. Theorem 7.1, however, requires us to correct from close to $1/2$ fraction of errors.

In Section 2 we remarked that it is possible to do *list*-decoding even after almost a $1/2$ fraction of errors occur. So we need a definition of *sub-linear time list decoding algorithm*. The definition is too technical to be given here. It was formulated, for a different application, in [5]. A reasonably simple sub-linear time list-decoding algorithm giving a new proof of Theorem 7.1 is presented in [68]. The coding-theoretic proof is considerably simpler than the original one.

The connection between error-correcting and hardness amplification also goes in the other direction: it is possible to view the techniques of [6], [27], [31] as defining list-decodable codes with sub-linear time decoding algorithm. This reverse connection has been used to transfer known coding theoretic impossibility results to the setting of amplification of hardness.

Recall that if we want to correct from $1/2 - \varepsilon$ errors, then unique decoding is impossible. Codes that are $(L, 1/2 - \varepsilon)$-list decodable exist, but it is possible to prove that for such codes we need $L = \Omega(\varepsilon^{-2})$. In our proof [68] of Theorem 6.2, this is not a problem because when we realize the decoding algorithm as a circuit we can "hard-wire" into the circuit the correct choice from the list. Suppose, however, that we want to prove a version of Theorem 6.2 where "algorithm of running time $K$" replaces "circuits of size $K$." Then such a theorem would not follow from [68]: if we try to follow the proof we see that from a good-on-average algorithm for $L' \cap \{0,1\}^{n'}$ we can only construct a *list of algorithms* such that one of them computes $L \cap \{0,1\}^n$ correctly, and it is not clear how to choose one algorithm out of this list.[12] This problem is solved in [74], where we do prove a version of Theorem 6.2 with "probabilistic

[12] This difficulty is discussed in [74].

algorithm" in place of "circuit."

Viola [81] proves that error-correcting codes cannot be computed in certain very low complexity classes, and this means that the exponentially big error-correcting code computations occurring in [68] must add a very strong complexity overhead. This means that coding-theoretic techniques cannot be used to prove a version of Theorem 6.2 where "computable in time $2^{O(n)}$" is replaced by "computable in NP." Indeed, it remains a fundamental open question whether a theorem showing equivalence of worst-case complexity and average-case complexity in NP can be proved. Results of [19], [10] show that this is unlikely.

Impagliazzo [28] wonders about a *positive* use of the fact that amplification of hardness results imply error-correcting codes, and whether the techniques of [6], [27], [31] would lead to practical error-correcting codes. We explore this question in [71], focusing on an optimization of the techniques of [27], but our results are far from being competitive with known constructions and algorithms of list-decodable codes. On the other hand, our work in refining the techniques of [27], while not successful in deriving good coding-theoretic applications, has led to interesting applications within complexity theory [71], [73].

## 8. Extractors and pseudorandom generators

We now come to what is perhaps the most surprising result of this survey, the fact that (the proofs of) Theorems 6.1 and 6.2 directly lead to *unconditional* constructions of extractors.

First, let us give a very high-level description of the pseudorandom generator construction that follows from Theorems 6.1 and 6.2.

Let $L$ be the set of exponential circuit complexity as in the assumption of Theorem 6.2, and let $m$ be a parameter such that we want to construct a generator $G_m\colon \{0,1\}^{O(\log m)} \to \{0,1\}^m$ whose output is $(m^2, 1/m)$-pseudorandom. First, we define $\ell = O(\log m)$ such that $L \cap \{0,1\}^\ell$ has circuit complexity at least $m^c$, for a certain absolute constant $c$. Then we define our generator as $G_m(z) = IW_m(L \cap \{0,1\}^\ell, z)$, where $IW_m(S, z)$ is a procedure that takes as input a set $S \subseteq \{0,1\}^\ell$ and a string $z \in \{0,1\}^{O(\log m)}$, outputs a string in $\{0,1\}^m$, and is such that if $S$ has circuit complexity at least $m^c$ then $IW_m(S, U_{O(\log m)})$ is $(m^2, 1/m)$-pseudorandom. Proving that $IW_m(\cdot,\cdot)$ has this property is of course quite complicated, but the general outline is as follows. As usual we proceed by contradiction and start from a statistical test $T$ of circuit complexity at most $m^2$ such that, supposedly,

$$|\,\mathbb{P}[U_m \in T] - \mathbb{P}[IW_m(S, U_{O(\log m)}) \in T]| > \tfrac{1}{m}.$$

Then we modify the circuit for $T$ and build a new circuit for $S$ of size $< m^c$, thus contradicting the hypothesis.

The analysis, indeed, proves a more general result. We will need some additional definitions before stating this more general result. For sets $T \subseteq \{0,1\}^m$ and $S \subseteq$

$\{0,1\}^\ell$, we say that $S$ has a *circuit with $T$-gates of size $K$* if there is a sequence of sets $S_1, \ldots, S_m$ such that $S_m = S$, and each $S_j$ is either a set of the form $P_{i,n}$, or it is the complement of a set $S_h$ $h < j$, or it is the union or the intersection of two sets $S_h$, $S_{h'}$ with $h, h' < j$, or it is defined as

$$S_j := \{a \in \{0,1\}^\ell : (1_{S_{h_1}}(a), \ldots, 1_{S_{h_m}}(a)) \in T\}$$

for some $h_1, \ldots, h_m < j$. It is not hard to show that if $S$ has a circuit with $T$-gates of size $K_1$, and $T$ has a regular circuit of size $K_2$, then $S$ has a regular circuit of size at most $K_1 \cdot K_2$. With these definitions in place we can be more specific about the analysis in [44], [31]: the analysis shows that if $S \subseteq \{0,1\}^\ell$ and $T \subseteq \{0,1\}^m$ are two *arbitrary* sets such that

$$|\,\mathbb{P}[U_m \in T] - \mathbb{P}[IW_m(S, U_{O(\log m)}) \in T]| > \tfrac{1}{m}$$

then there is a circuit with $T$-gates for $S$ of size $< m^{c-2}$. (Note that this implies our previous statement.)

Here is the main idea in [70]: suppose that we have access to a weak random source, that is, a random variable $X$ taking values in $\{0,1\}^n$ and having min-entropy at least $k$. Suppose that $n = 2^\ell$. Then we can, equivalently, regard $X$ as being distributed over $\mathcal{P}(\{0,1\}^\ell)$, the set of all subsets of $\{0,1\}^\ell$. What can we say about the distribution of $IW_m(X, U_{O(\log m)})$? We claim that, if $k$ is large enough, the distribution $IW_m(X, U_{O(\log m)})$ is close in statistical distance to the uniform distribution; in other words, $IW_m(\cdot,\cdot)$ is an *extractor*.

Let us see how to prove this by contradiction. Let $T$ be a statistical test such that

$$|\,\mathbb{P}[U_m \in T] - \mathbb{P}[IW_m(X, U_{O(\log m)}) \in T]| > \tfrac{1}{m}$$

and call a set $S \in \mathcal{P}(\{0,1\}^\ell)$ *bad* if $|\,\mathbb{P}[U_m \in T] - \mathbb{P}[IW_m(S, U_{O(\log m)}) \in T]| > \frac{2}{m}$.

Let $B$ be the set of all bad sets. Then, by Markov's inequality, $\mathbb{P}[X \in B] \geq \frac{2}{m}$, and since $X$ has min-entropy $k$ we have $|B| \geq 2^{k-\log m-1}$. On the other hand, if $S$ is bad, then there is a circuit with $T$-gates of size at most $m^{c-2}$ that computes $S$. The number of such circuits is at most $2^{O(m^{c-1})}$, and so $|B| \leq 2^{O(m^{c-1})}$. So if $k \geq c' m^{c-1}$, where $c$, $c'$ are absolute constants, we reach a contradiction. Thus, $\|IW_m(X, U_{O(\log m)}) - U_m\|_{\mathrm{SD}} \leq \frac{1}{m}$.

If we look more closely at how $IW_m(S, z)$ is defined, we see that (especially if we use the proof of Theorem 6.2 in [68]) it can be seen as $IW_m(S, z) := NW_m(C(S), z)$, where $C$ is an error-correcting code and $NW_m$ is the relatively simple pseudorandom generator construction of Nisan and Wigderson. For the application to derandomization, it is important that $C$ be a "sub-linear time list-decodable" error-correcting code. However, in order for our argument about randomness extraction to work, it is sufficient that $C$ be an arbitrary list-decodable code, and not even a polynomial time list-decoding algorithm is needed. This means that one can get extractors by using standard error-correcting codes and the simple Nisan–Wigderson generator. The

resulting construction is described and analysed in [70] in about two pages and, at the time, it was the best known extractor construction, improving over very technical previous work.

What makes these calculations work is the intuition that the proofs of Theorems 6.1 and 6.2 prove more than the intended statement. In particular, the proof works if we replace "circuit complexity" with "description complexity" which we exploited in the previous argument. See [70] for further discussion of this point.

The connection with pseudorandomness and the general idea of analysing an extractor by finding short descriptions of the output of the source based on a hypothetical statistical test (the so-called "reconstruction method" to analyse extractors) has led to remarkable advances in extractor constructions in the past five years, together with other ideas. The best distillation of the reconstruction method is in [58], providing a near-optimal and simple construction of extractors.[13] The extractor motivation has also led to improvements in pseudorandom generator constructions, see [58], [75]. Currently, the best known extractor construction [40] uses the notion of "condenser" introduced in [69], [54] and a combination of several components, one of which is analysed with the reconstruction method. The extractors of Lu et al. [40] is almost best possible.

## 9. Conclusions

We have discussed how, starting from worst-case complexity assumptions, it is possible to construct very strong pseudorandom generators and derive conditional derandomization results for *all* probabilistic algorithms.

What about *proving* circuit lower bounds and deriving unconditional derandomization results? The results of Razborov and Rudich [52] show that a significant departure from current techniques will be required to prove such lower bounds. What about deriving derandomization results *without* proving lower bounds? Impagliazzo, Kabanets and Wigderson [29] prove that any general derandomization result implies a circuit lower bound.[14]

Short of proving such elusive circuit lower bounds, we should test the prediction of the theory and look for polynomial time deterministic versions of known probabilistic polynomial time algorithms. The four most important probabilistic algorithms (or collections of algorithms) are: primality testing, graph connectivity using random walks, polynomial identity testing, and algorithms for approximate counting. Primality testing and graph connectivity using random walks have been derandomized [1], [53]. Kabanets and Impagliazzo [34] prove that any derandomized polynomial identity testing algorithms implies circuit lower bounds.[15]

[13] The construction is simple but the analysis is quite non-trivial.

[14] Here is what we mean by "general derandomization": if $f$ is a function and $A$ is randomized algorithm that with high probability achieves a good approximation of $f$, then there is a deterministic algorithm that achieves a good approximation of $f$ and whose running time is polynomial in the running time of $A$.

The possibility of derandomizing approximate counting algorithms with current techniques is quite open. Here is perhaps the simplest question: given an $n$-variable boolean formula in disjunctive normal form and $\varepsilon > 0$, compute in time polynomial in the size of the formula and in $1/\varepsilon$ an approximation to the number of satisfying assignments up to an additive error $\leq 2^n\varepsilon$. See [42] for a nearly polynomial time deterministic algorithm for this problem.

The construction of an optimal (seeded) extractor with parameters matching the known lower bounds remains an elusive open question. It would also be interesting to match the parameters of [40] with a simpler construction.

There has been very exciting recent progress towards constructing good seedless extractors for independent sources, and for the related problem of constructing bipartite Ramsey graphs [8], [11]. The broader area of seedless extractor constructions for general classes of distributions has seen much recent progress. In the long run, we would expect this research to define simple and powerful seedless extractors working for a wide and natural class of distributions. Such extractors would be very useful in practice, giving a principled approach to the production of random bits for cryptographic applications.

## References

[1] Agrawal, Manindra, Kayal, Neeraj, and Saxena, Nitin, PRIMES is in P. *Ann. of Math.* **160** (2) (2004), 781–793.

[2] Aleliunas, Romas, Karp, Richard M., Lipton, Richard J., Lovász, László, and Rackoff, Charles, Random walks, universal traversal sequences, and the complexity of maze problems. In *Proceedings of the 20th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1979, 218–223.

[3] Alon, Noga, and Capalbo, Michael R., Explicit unique-neighbor expanders. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2002, 73–79.

[4] Alon, Noga, and Spencer, Joel, *The Probabilistic Method*. Wiley-Intersci. Ser. Discrete Math. Optim., John Wiley and Sons, New York 2000.

[5] Arora, Sanjeev, and Sudan, Madhu, Improved low degree testing and its applications. *Combinatorica* **23** (3) (2003), 365–426.

[6] Babai, László, Fortnow, Lance, Nisan, Noam, and Wigderson, Avi, BPP has subexponential time simulations unless EXPTIME has publishable proofs. *Comput. Complexity* **3** (4) (1993), 307–318.

[7] Barak, Boaz, Impagliazzo, Russell, and Wigderson, Avi, Extracting randomness using few independent sources. In *Proceedings of the 45th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2004, 384–393.

[15] Fortunately, these are not of the kind ruled out by [52], so there is some hope. Indeed Raz [49], [50] has recently proved lower bounds that are weaker than, but in the spirit of, what is needed to derandomize polynomial identity testing.

[8] Barak, Boaz, Kindler, Guy, Shaltiel, Ronen, Sudakov, Benny, and Wigderson, Avi, Simulating independence: new constructions of condensers, Ramsey graphs, dispersers, and extractors. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 1–10.

[9] Blum, Manuel, and Micali, Silvio, How to generate cryptographically strong sequences of pseudorandom bits. *SIAM J. Comput.* **13** (4) (1984), 850–864.

[10] Bogdanov, Andrej, and Trevisan, Luca, On wost-case to average-case reductions for NP problems. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2003, 308–317.

[11] Bourgain, Jean, More on the sum-product phenomenon in prime fields and its applications. *Int. J. Number Theory* **1** (1) (2005), 1–32.

[12] Bourgain, Jean, Katz, Nets, and Tao, Terence, A sum-product estimate for finite fields, and applications. *Geom. Funct. Anal.* **14** (2004), 27–57.

[13] Capalbo, Michael R., Reingold, Omer, Vadhan, Salil P., and Wigderson, Avi, Randomness conductors and constant-degree lossless expanders. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, ACM Press, New York 2002, 659–668.

[14] Chor, Benny, and Goldreich, Oded, Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.* **17** (2) (1988), 230–261.

[15] Chung, Fan R. K., Graham, Ronald L., and Wilson, Richard M., Quasi-random graphs. *Combinatorica* **9** (4) (1989), 345–362.

[16] Cohen, Aviad, and Wigderson, Avi, Dispersers, deterministic amplification, and weak random sources. In *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1989, 14–19.

[17] Elias, Peter, List decoding for noisy channels. Technical Report 335, Research Laboratory of Electronics, MIT, 1957.

[18] Erdős, Paul, Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.* **53** (1947), 292–294.

[19] Feigenbaum, Joan, and Fortnow, Lance, Random-self-reducibility of complete sets. *SIAM J. Comput.* **22** (1993), 994–1005.

[20] Frankl, Peter, and Wilson, Richard M., Intersection theorems with geometric consequences. *Combinatorica* **1** (4) (1981), 357–368.

[21] Goldreich, Oded, *Foundations of Cryptography.* Volume 1, Cambridge University Press, Cambridge 2001.

[22] Goldwasser, Shafi, and Micali, Silvio, Probabilistic encryption. *J. Comput. System Sci.* **28** (2) (1984), 270–299.

[23] Guruswami, Venkatesan, List Decoding of Error-Correcting Codes. PhD thesis, MIT, 2001.

[24] Guruswami, Venkatesan, and Rudra, Atri, Explicit capacity-achieving list-decodable codes. Technical Report TR05-133, Electronic Colloquium on Computational Complexity, 2005.

[25] Hamming, Richard, Error detecting and error correcting codes. *Bell System Tech. J.* **29** (1950), 147–160.

[26] Håstad, Johan, Impagliazzo, Russell, Levin, Leonid, and Luby, Michael, A pseudorandom generator from any one-way function. *SIAM J. Comput.* **28** (4) (1999), 1364–1396.

[27] Impagliazzo, Russell, Hard-core distributions for somewhat hard problems. In *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1995, 538–545.

[28] Impagliazzo, Russell, Hardness as randomness: a survey of universal derandomization. *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing 2002, 659–672.

[29] Impagliazzo, Russell, Kabanets, Valentine, and Wigderson, Avi, In search of an easy witness: exponential time vs. probabilistic polynomial time. *J. Comput. System Sci.* **65** (4) (2002), 672–694.

[30] Impagliazzo, Russell, Nisan, Noam, and Wigderson, Avi, Pseudorandomness for network algorithms. In *Proceedings of the 26th ACM Symposium on Theory of Computing*, ACM Press, New York 1994, 356–364.

[31] Impagliazzo, Russell, and Wigderson, Avi, $P = BPP$ unless $E$ has sub-exponential circuits. In *Proceedings of the 29th ACM Symposium on Theory of Computing*, ACM Press, New York 1997, 220–229.

[32] Iwama, Kazuo, and Morizumi, Hiroki, An explicit lower bound of $5n - o(n)$ for boolean circuits. In *Proceedings of the 27th Symposium on Mathematical Foundations of Computer Science*, Lecture Notes in Comput. Sci. 2420, Springer-Verlag, London 2002, 353–364.

[33] Jerrum, Mark, Sinclair, Alistair, and Vigoda, Eric, A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. ACM* **51** (4) (2004), 671–697.

[34] Kabanets, Valentine, and Impagliazzo, Russell, Derandomizing polynomial identity tests means proving circuit lower bounds. *Comput. Complexity* **13** (1–2) (2004), 1–46.

[35] Klivans, Adam, and van Melkebeek, Dieter, Graph nonisomorphism has subexponential size proofs unless the polynomial-time hierarchy collapses. *SIAM J. Comput.* **31** (5) (2002), 1501–1526.

[36] Konyagin, Sergei, A sum-product estimate in fields of prime order. math.NT/0304217, 2003.

[37] Krivelevich, Michael, and Sudakov, Benny, Pseudo-random graphs. Preprint, 2005.

[38] Lachish, Oded, and Raz, Ran, Explicit lower bound of $4.5n - o(n)$ for boolean circuits. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*, ACM Press, New York 2001, 399–408.

[39] Lu, Chi-Jen, Encryption against storage-bounded adversaries from on-line strong extractors. *J. Cryptology* **17** (1) (2004), 27–42.

[40] Lu, Chi-Jen, Reingold, Omer, Vadhana, Salil P., and Wigderson, Avi, Extractors: optimal up to constant factors. In *Proceedings of the 35th ACM Symposium on Theory of Computing*, ACM Press, New York, 602–611.

[41] Lubotzky, Alexander, Phillips, R., and Sarnak, Peter, Ramanujan graphs. *Combinatorica* **8** (1988), 261–277.

[42] Luby, Michael, and Velickovic, Boban, On deterministic approximation of DNF. *Algorithmica* **16** (4/5) (1996), 415–433.

[43] Miltersen, Peter B., and Vinodchandran, N. V., Derandomizing Arthur-Merlin games using hitting sets. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1999, 71–80.

[44] Nisan, Noam, and Wigderson, Avi, Hardness vs randomness. *J. Comput. System Sci.* **49** (1994), 149–167.

[45] Nisan, Noam, and Zuckerman, David, Randomness is linear in space. *J. Comput. System Sci.* **52** (1) (1996), 43–52.

[46] Parvaresh, Farzad, and Vardy, Alexander, Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2005, 285–294.

[47] Rabin, Michael, Probabilistic algorithm for testing primality. *J. Number Theory* **12** (1980), 128–138.

[48] Rao, Anup, Extractors for a constant number of polynomially small min-entropy independent sources. In *Proceedings of the 38th ACM Symposium on Theory of Computing*, ACM Press, New York 2006, 497–506.

[49] Raz, Ran, Multi-linear formulas for permanent and determinant are of super-polynomial size. In *Proceedings of the 36th ACM Symposium on Theory of Computing*, ACM Press, New York 2004, 633–641.

[50] Raz, Ran, Multilinear-$NC_1 \neq$ multilinear-$NC_2$. In *Proceedings of the 45th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2004, 344–351.

[51] Raz, Ran, and Reingold, Omer, On recycling randomness in space bounded computation. In *Proceedings of the 31st ACM Symposium on Theory of Computing*, ACM Press, New York 1999, 159–168.

[52] Razborov, Alexander A., and Rudich, Steven, Natural proofs. *J. Comput. System Sci.* **55** (1) (1997), 24–35.

[53] Reingold, Omer, Undirected ST-connectivity in log-space. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 376–385.

[54] Reingold, Omer, Shaltiel, Ronen, and Wigderson, Avi, Extracting randomness by repeated condensing. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2000, 22–31.

[55] Santha, Miklos, and Vazirani, Umesh, Generating quasi-random sequences from slightly random sources. *J. Comput. System Sci.* **33** (1986), 75–87.

[56] Schwartz, Jacob T., Fast probabilistic algorithms for verification of polynomial identities. *J. ACM* **27** (1980), 701–717.

[57] Shaltiel, Ronen, Recent developments in explicit constructions of extractors. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **77** (2002), 67–95.

[58] Shaltiel, Ronen, and Umans, Christopher, Simple extractors for all min-entropies and a new pseudorandom generator. *J. ACM* **52** (2) (2005), 172–216.

[59] Shannon, Claude, A mathematical theory of communications. *Bell System Tech. J.* **27** (1948), 379–423, 623–656, 1948.

[60] Shannon, Claude, The synthesis of two-terminal switching circuits. *Bell System Tech. J.* **28** (1949), 59–98.

[61] Sinclair, Alistair, and Jerrum, Mark, Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.* **82** (1) (1989), 93–133.

[62] Sipser, Michael, Borel sets and circuit complexity. In *Proceedings of the 15th ACM Symposium on Theory of Computing*, ACM Press, New York 1983, 61–69.

[63] Sipser, Michael, A topological view of some problems in complexity theory. In *Proceedings of the Symposium on Mathematical Foundations of Computer Science*, IEEE, New York 1984, 567–572.

[64] Sipser, Michael, *Introduction to the Theory of Computation*. PWS Publishing Co., Boston, MA, 1997.

[65] Solovay, Robert, and Strassen, Volker, A fast Monte-Carlo test for primality. *SIAM J. Comput.* **6** (1) (1977), 84–85.

[66] Sudan, Madhu, Decoding of Reed-Solomon codes beyond the error-correction bound. *J. Complexity* **13** (1) (1997), 180–193.

[67] Sudan, Madhu, List decoding: Algorithms and applications. *SIGACT News* **31** (1) (2000), 16–27.

[68] Sudan, Madhu, Trevisan, Luca, and Vadhan, Salil, Pseudorandom generators without the XOR lemma. *J. Comput. System Sci.* 62 (2) (2001), 236–266,.

[69] Ta-Shma, Amnon, Umans, Christopher, and Zuckerman, David, Loss-less condensers, unbalanced expanders, and extractors. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*, ACM Press, New York 2001, 143–152.

[70] Trevisan, Luca, Extractors and pseudorandom generators. *J. ACM* **48** (4) (2001), 860–879.

[71] Trevisan, Luca, List-decoding using the XOR Lemma. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2003, 126–135.

[72] Trevisan, Luca, Some applications of coding theory in computational complexity. *Quad. Mat.* **13** (2004), 347–424.

[73] Trevisan, Luca, On uniform amplification of hardness in NP. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, ACM Press, New York 2005, 31–38.

[74] Trevisan, Luca, and Vadhan, Salil, Pseudorandomness and average-case complexity via uniform reductions. In *Proceedings of the 17th IEEE Conference on Computational Complexity*, IEEE, New York 2002, 129–138.

[75] Umans, Christopher, Pseudo-random generators for all hardnesses. *J. Comput. System Sci.* **67** (2) (2003), 419–440.

[76] Vadhan, Salil, Randomness extractors and their many guises. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science*, IEEE, New York 2002, 9–10.

[77] Vadhan, Salil P., Constructing locally computable extractors and cryptosystems in the bounded-storage model. *J. Cryptology* **17** (1) (2004), 43–77.

[78] Vazirani, Umesh, Randomness, Adversaries and Computation. PhD thesis, University of California, Berkeley, 1986.

[79] Vazirani, Umesh, and Vazirani, Vijay, Random polynomial time is equal to slightly random polynomial time. In *Proceedings of the 26th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1985, 417–428.

[80] Vazirani, Umesh V., Strong communication complexity or generating quasirandom sequences form two communicating semi-random sources. *Combinatorica* **7** (4) (1987), 375–392.

[81] Viola, Emanuele, The complexity of constructing pseudorandom generators from hard functions. *Comput. Complexity* **13** (3–4) (2004), 147–188.

[82] von Neumann, John, Various techniques used in connection with random digits. *J. Res. Nat. Bur. Standards App. Math. Ser.* **12** (1951), 36–38.

[83] Wigderson, Avi, and Zuckerman, David, Expanders that beat the eigenvalue bound: Explicit construction and applications. *Combinatorica* **19** (1) (1999), 125–138.

[84] Yao, Andrew C., Theory and applications of trapdoor functions. In *Proceedings of the 23th IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1982, 80–91.

[85] Zippel, Richard, Probabilistic algorithms for sparse polynomials. In *Symbolic and algebraic computation* (ed. by Edward W. Ng), Lecture Notes in Comput. Sci. 72, Springer-Verlag, Berlin 1979, 216–226.

[86] Zuckerman, David, General weak random sources. In *Proceedings of the 31st IEEE Symposium on Foundations of Computer Science*, IEEE, New York 1990, 534–543.

[87] Zuckerman, David, Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the 38th ACM Symposium on Theory of Computing*, ACM Press, New York 2006, 681–690.

Computer Science Division, U.C. Berkeley, 679 Soda Hall, Berkeley, CA 94720-1776, U.S.A.

E-mail: luca@cs.berkeley.edu

# Least-squares finite element methods

Pavel Bochev* and Max Gunzburger†

**Abstract.** Least-squares finite element methods are an attractive class of methods for the numerical solution of partial differential equations. They are motivated by the desire to recover, in general settings, the advantageous features of Rayleigh–Ritz methods such as the avoidance of discrete compatibility conditions and the production of symmetric and positive definite discrete systems. The methods are based on the minimization of convex functionals that are constructed from equation residuals. This paper focuses on theoretical and practical aspects of least-square finite element methods and includes discussions of what issues enter into their construction, analysis, and performance. It also includes a discussion of some open problems.

**Mathematics Subject Classification (2000).** 65N30, 65N99, 65N15.

**Keywords.** Least squares, finite element methods, compatible discretizations.

## 1. Introduction

Finite element methods (FEMs) for the approximate numerical solution of partial differential equations (PDEs) were first developed and analyzed for problems in linear elasticity and other settings for which solutions can be characterized as (unconstrained) minimizers of convex, quadratic functionals on infinite-dimensional Hilbert spaces [46]. A Rayleigh–Ritz approximation of such solutions is defined by minimizing the functional over a family of finite-dimensional subspaces. An FEM results when these spaces consist of piecewise polynomial functions defined with respect to a family of grids. When applied to problems such as linear elasticity or the Poisson equation, the Rayleigh–Ritz setting gives rise to FEMs with several advantageous features that led to their great success and popularity:

1. general regions and boundary conditions are relatively easy to treat in a systematic manner;
2. the conformity[1] of the finite element spaces suffices to guarantee the stability

*Supported by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research, and performed at Sandia National Labs, a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

†Supported in part by CSRI, Sandia National Laboratories, under contract 18407 and by the National Science Foundation under grant number DMS-0240049.

[1]An approximating space is referred to as being *conforming* if it is a subspace of the underlying infinite-dimensional Hilbert space.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

and optimal accuracy[2] of the approximate solutions;

3. all variables can be approximated using a single type of finite element space, e.g., the same degree piecewise polynomials defined with respect to a same grid;
4. the resulting linear systems are
   a) sparse; b) symmetric; c) positive definite.

The success of FEMs in the Rayleigh–Ritz setting quickly led both engineers and mathematicians to apply and analyze FEMs in other settings, motivated by the fact that properties 1 and 4a are retained for all FEMs.[3] For example, *mixed* FEMs arose from minimization problems constrained by PDEs such as the Stokes problem; the Lagrange multiplier rule was applied to enforce the constraints, resulting in saddle-point problems [19]. In this setting, the only other property retained from the Rayleigh–Ritz setting is 4b. More generally, *Galerkin* FEMs can, in principle, be defined for any PDE by forcing the residual of the PDE (posed in a weak, variational formulation) to be orthogonal to the finite element subspace [3]. In this general setting, one usually loses all the features of the Rayleigh–Ritz setting other than 1 and 4a. Using the same formalisms, Galerkin FEMs were even applied to *nonlinear* problems such as the Navier–Stokes equations [34]. It is a testament to the importance of advantage 1 that despite the loss of other advantages, mixed and Galerkin FEMs are in widespread use and have also been extensively analyzed.[4]

Not surprisingly, despite the success of mixed and Galerkin FEMs in general settings, there has been substantial interest and effort devoted to developing finite element approaches that recover at least some of the advantages of the Rayleigh–Ritz setting. Notable among these efforts have been penalty and stabilized FEMs, e.g., for the Stokes problem, stabilized FEMs [4]–[5], [6], [12], [15], [29], [30], [36], [37], [44] recover advantages 2 and 3 but fail to recover advantage 4c and often lose advantage 4b.

Least-squares finite element methods (LSFEMs) can be viewed as another attempt at retaining the advantages of the Rayleigh–Ritz setting even for much more general problems. In fact, they offer the possibility of, in principle, retaining *all* of the advantages of that setting for practically any PDE problem. In §2, we show how this is possible. However, this is not the whole story. Any FEM, including an LSFEM, must also meet additional practicality criteria:

A. bases for conforming subspaces are easily constructed;

B. linear systems are easily assembled;

C. linear systems are relatively well conditioned.

[2] An approximate solution is referred to as being *optimally accurate* if the corresponding error is bounded by a constant times the error of the best approximation.

[3] These properties follow from the way finite element spaces are constructed, e.g., based on grids and choosing basis functions of compact support.

[4] It should be noted that in the general settings for which FEMs lose many of the advantages they possess in the Rayleigh–Ritz setting, they do not suffer from any disadvantages compared to other discretization methods such as finite difference, finite volume, and spectral methods.

In judging whether or not an LSFEM meets theses criteria, we will measure them up against Galerkin FEMs for the Poisson equation; in particular, we will ask the questions: can we use standard, piecewise polynomial spaces that are merely continuous and for which bases are easily constructed? can the assembly of the linear systems be accomplished by merely applying quadrature rules to integrals? and, are the condition numbers of the linear systems of[5] $O(h^{-2})$? Unfortunately, naively defined LSFEMs often fail to meet one or more of the practicality criteria.

LSFEMs possess two additional advantageous features that other FEMs, even in the Rayleigh–Ritz setting, do not possess. First, least-square functionals provide an easily computable residual error indicator that can be used for adapting grids. Second, the treatment of general boundary conditions, including nonhomogeneous ones, is greatly facilitated because boundary condition residuals can be incorporated into the least-square functional.

## 2. The most straightforward LSFEM

Let $\Omega$ denote a bounded domain in $\Re^d$, $d = 2$ or 3, with boundary $\Gamma$. Consider the problem

$$\mathcal{L}u = f \text{ in } \Omega \quad \text{and} \quad \mathcal{R}u = g \text{ on } \Gamma, \tag{1}$$

where $\mathcal{L}$ is a linear differential operator and $\mathcal{R}$ is a linear boundary operator. We assume that the problem (1) is well posed so that there exists a solution Hilbert space $S$, data Hilbert spaces $H_\Omega$ and $H_\Gamma$, and positive constants $\alpha_1$ and $\alpha_2$ such that

$$\alpha_1 \|u\|_S^2 \le \|\mathcal{L}u\|_{H_\Omega}^2 + \|\mathcal{R}u\|_{H_\Gamma}^2 \le \alpha_2 \|u\|_S^2 \quad \text{for all } u \in S. \tag{2}$$

Then consider the least-squares functional[6]

$$J(u; f, g) = \|\mathcal{L}u - f\|_{H_\Omega}^2 + \|\mathcal{R}u - g\|_{H_\Gamma}^2 \tag{3}$$

and the unconstrained minimization problem

$$\min_{u \in S} J(u; f, g). \tag{4}$$

Note that the functional (3) measures the residuals of the components of the system (1) using the data space norms $H_\Omega$ and $H_\Gamma$ and the minimization problem (4) seeks a solution in the solution space $S$ for which (2) is satisfied. It is clear that the problems (1) and (4) are equivalent in the sense that $u \in S$ is a solution of (4) if and only if it is also a solution, perhaps in a generalized sense, of (1).

An LSFEM can be defined by choosing a family of finite element subspaces $S^h \subset S$ parameterized by $h$ tending to zero and then restricting the minimization problem (4)

[5] Usually, $h$ is a measure of the size of the grid used in the construction of the finite element space.

[6] A least-squares functional may be viewed as an "artificial" energy that plays the same role for LSFEMs as a bona fide physically energy plays for Rayleigh–Ritz FEMs.

to the subspaces. Thus, the LSFEM approximation $u^h \in S^h$ to the solution $u \in S$ of (1) or (4) is the solution of the problem

$$\min_{u^h \in S^h} J(u^h; f, g). \tag{5}$$

The Euler–Lagrange equations corresponding to the minimization problems (4) and (5) are given by

$$\textit{seek } u \in S \quad \textit{such that} \quad B(u, v) = F(v) \quad \textit{for all } v \in S, \tag{6}$$

$$\textit{seek } u^h \in S^h \quad \textit{such that} \quad B(u^h, v^h) = F(v^h) \quad \textit{for all } v^h \in S^h, \tag{7}$$

respectively, where for all $u, v \in S$,

$$B(u, v) = (\mathcal{L}v, \mathcal{L}u)_{H_\Omega} + (\mathcal{R}v, \mathcal{R}u)_{H_\Gamma} \quad \text{and} \quad F(v) = (\mathcal{L}v, f)_{H_\Omega} + (\mathcal{R}v, g)_{H_\Gamma}. \tag{8}$$

If we choose a basis $\{U_j\}_{j=1}^J$, where $J = \dim(S^h)$, then we have that $u^h = \sum_{j=1}^J c_j U_j$ for some constants $\{c_j\}_{j=1}^J$ and then the discretized problem (7) is equivalent to the linear system

$$\mathbb{K}\boldsymbol{c} = \boldsymbol{f}, \tag{9}$$

where the elements of the matrix $\mathbb{K} \in \Re^{J \times J}$ and the vectors $\boldsymbol{f} \in \Re^J$ and $\boldsymbol{c} \in \Re^J$ are given, for $i, j = 1, \ldots, J$, by $\boldsymbol{c}_j = c_j$,

$$\mathbb{K}_{ij} = (\mathcal{L}U_i, \mathcal{L}U_j)_{H_\Omega} + (\mathcal{R}U_i, \mathcal{R}U_j)_{H_\Gamma}, \quad \text{and} \quad \boldsymbol{f}_i = (\mathcal{L}U_i, f)_{H_\Omega} + (\mathcal{R}U_i, g)_{H_\Gamma}.$$

The results of the following theorem follow directly from (2).

**Theorem 2.1.** *Assume that* (2) *holds and that* $S^h \subset S$. *Then,*

*– the bilinear form* $B(\cdot, \cdot)$ *defined in* (8) *is continuous, symmetric, and coercive;*

*– the linear functional* $F(\cdot)$ *defined in* (8) *is continuous;*

*– the problem* (6) *has a unique solution* $u \in S$ *that is also the unique solution of the minimization problem* (4)*;*

*– the problem* (7) *has a unique solution* $u^h \in S^h$ *that is also the unique solution of the minimization problem* (5)*;*

*– for some constant* $C > 0$, *we have that* $\|u\|_S \leq C(\|f\|_{H_\Omega} + \|g\|_{H_\Gamma})$ *and* $\|u^h\|_S \leq C(\|f\|_{H_\Omega} + \|g\|_{H_\Gamma})$;

*– for some constant* $C > 0$, $u$ *and* $u^h$ *satisfy the error estimate*

$$\|u - u^h\|_S \leq C \inf_{v^h \in S^h} \|u - v^h\|_S; \tag{10}$$

*– the matrix* $\mathbb{K}$ *of* (9) *is symmetric and positive definite.*

Note that it is not assumed that the system (1) is self-adjoint or positive as it would have to be in the Rayleigh–Ritz setting; it is only assumed that it is well posed. Despite the generality of the system (1), the LSFEM based on (5) recovers *all* desirable

features of FEMs in the Rayleigh–Ritz setting. Note that (10) shows that least-squares finite element approximations are optimally accurate with respect to solution norm $\|\cdot\|_S$ for which the system (1) is well posed.

In defining the least-squares principle (4), we have not restricted the spaces $S$ and $S^h$ to satisfy the boundary conditions. Instead, we have included the residual $\mathcal{R}u - g$ of the boundary condition in the functional $J(\cdot\,;\cdot,\cdot)$ defined in (3). Thus, we see that LSFEMs possess a desirable feature that is absent even from standard FEMs in the Rayleigh–Ritz setting: the imposition of boundary conditions can be effected through the functional and need not be imposed on the finite element spaces.[7] Notwithstanding this advantage, one can impose essential boundary conditions on the space $S$ in which case all terms in (2)–(8) involving the boundary condition are omitted and we also set $H_\Omega = H$. Note also that since

$$\begin{aligned} J(u_h;\, f, g) &= \|\mathcal{L}u_h - f\|^2_{H_\Omega} + \|\mathcal{R}u_h - g\|^2_{H_\Gamma} \\ &= B(u_h, u_h) - 2F(u_h) + (f,\, f)_{H_\Omega} + (g, g)_{H_\Gamma}, \end{aligned}$$

the least-square functional $J(u_h;\, f, g)$ provides a *computable* indicator for the residual error in the LSFEM approximation $u^h$. Such indicators are in widespread used for grid adaption.

The problems (6) and (7) display the *normal equation* form typical of least-squares systems; see (8). It is important to note that since $\mathcal{L}$ is a differential operator, (6) involves a higher-order differential operator. We shall see that this observation has a profound effect on how practical LSFEMs are defined.

**2.1. The practicality of the straightforward LSFEM.** The complete recovery, in general settings, of all desirable features of the Rayleigh–Ritz setting is what makes LSFEMs intriguing and attractive. But, what about the practicality of the method defined by (5)? We explore this issue using examples.

**2.1.1. An impractical application of the straightforward LSFEM.** Consider the problem

$$-\Delta u = f \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \Gamma, \tag{11}$$

where we assume that $\Omega$ is either a convex, Lipschitz domain or that it has a smooth boundary. Of course, this is a problem which fits into the Rayleigh–Ritz framework so that there is no apparent need[8] to use any other type of FEM. However, let us proceed and use the LSFEM method anyway, and see what happens. Here we have that (2) holds with[9] $S = H^2(\Omega) \cap H^1_0(\Omega)$, $H = L^2(\Omega)$, and $\mathcal{L} = -\Delta$. We then have

[7] This advantage of LSFEM can be useful for imposing inhomogeneous boundary conditions, essential boundary conditions such as Dirichlet boundary conditions for second-order elliptic PDEs, and boundary conditions involving a particular component, e.g., the normal component, of a vector variable.

[8] Inhomogeneous Dirichlet boundary conditions provide a situation in which one might want to use LSFEMs even for the Poisson problem.

[9] We use standard Sobolev space notation throughout the paper. Also, in this and most of our examples, we will be imposing the boundary condition on the solutions space $S$.

that, for all $u, v \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$J(u; f) = \|\Delta u + f\|_0^2, \quad F(v) = \int_\Omega f \Delta v \, d\Omega, \quad \text{and} \quad B(u, v) = \int_\Omega \Delta v \, \Delta u \, d\Omega.$$

Note that minimizing the least-squares functional has turned the second-order Poisson problem into a fourth-order problem.

An LSFEM is defined by choosing a subspace $S^h \subset S = H^2(\Omega) \cap H_0^1(\Omega)$ and then posing the problem (7). It is well known that in this case, the finite element space $S^h$ has to consist of continuously differentiable functions; this requirement greatly complicates the construction of bases and the assembly of the matrix problem. Furthermore, it is also well known that the condition number of the matrix problem is $O(h^{-4})$ which should be contrasted with the $O(h^{-2})$ condition number obtained through a Rayleigh–Ritz discretization of the Poisson equation. Thus, for this problem, the straightforward LSFEM fails all three practicality tests.

Since it is also true that (2) holds with $S = H_0^1(\Omega)$ and $H = H^{-1}(\Omega)$, one could develop an LSFEM based on the functional $J(u; f) = \|\Delta u + f\|_{-1}$ and the solution space $S = H_0^1(\Omega)$. This approach would allow one to use a finite element space $S^h$ consisting of merely continuous functions so that bases may be easily constructed. Moreover, it can be shown that because of the use of the $H^{-1}(\Omega)$ inner product, the condition number of the resulting matrix system is $O(h^{-2})$ which is the same as for a Rayleigh–Ritz discretization. However, the $H^{-1}(\Omega)$ inner product is computed by inverting the Laplacian operator which leads to the loss of property 4a and also makes the assembly of the matrix problem more difficult. So, as it stands, the straightforward LSFEM remains impractical for the second-order Poisson problem.

**2.1.2. A practical application of the straightforward LSFEM.** Consider now the problem

$$-\nabla \cdot \boldsymbol{u} = f \ \text{ and } \ \nabla \times \boldsymbol{u} = \boldsymbol{g} \ \text{ in } \Omega \quad \text{and} \quad \boldsymbol{n} \cdot \boldsymbol{u} = 0 \ \text{ on } \Gamma. \tag{12}$$

Here $\boldsymbol{u} \in S = \boldsymbol{H}_n^1(\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) \,|\, \boldsymbol{n} \cdot \boldsymbol{v} = 0 \text{ on } \Gamma\}$ and $\{f, \boldsymbol{g}\} \in H = L_0^2(\Omega) \times \boldsymbol{L}_s^2(\Omega)$, where $L_0^2(\Omega) = \{f \in L^2(\Omega) \,|\, \int_\Omega f \, d\Omega = 0\}$, and $\boldsymbol{L}_s^2(\Omega) = \{\boldsymbol{g} \in \boldsymbol{L}^2(\Omega) \,|\, \nabla \cdot \boldsymbol{g} = 0 \text{ in } \Omega\}$. We then have that (2) holds so that we may define the least-squares functional

$$J(\boldsymbol{u}; f, \boldsymbol{g}) = \|\nabla \cdot \boldsymbol{u} + f\|_0^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{g}\|_0^2 \quad \text{for all } \boldsymbol{u} \in S = \boldsymbol{H}_n^1(\Omega) \tag{13}$$

that results in

$$B(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \big((\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{v}) + (\nabla \times \boldsymbol{u}) \cdot (\nabla \times \boldsymbol{v})\big) \, d\Omega \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in S = \boldsymbol{H}_n^1(\Omega)$$

and

$$F(\boldsymbol{v}) = \int_\Omega \big( - f \nabla \cdot \boldsymbol{v} + \boldsymbol{g} \cdot \nabla \times \boldsymbol{v}\big) \, d\Omega \quad \text{for all } \boldsymbol{v} \in S = \boldsymbol{H}_n^1(\Omega).$$

An LSFEM is defined by choosing a subspace $S^h \subset S = \boldsymbol{H}_n^1(\Omega)$ and then solving the problem (7).

The LSFEM based on the functional (13) not only recovers all the good properties of the Rayleigh–Ritz setting for the problem (12), but also satisfies all three practicality criteria. Since we merely require that $S^h \subset \boldsymbol{H}_n^1(\Omega)$, we can choose standard finite element spaces for which bases are easily constructed. Furthermore, since the functional (13) only involves $L^2(\Omega)$ inner products, the assembly of the matrix system is accomplished in a standard manner. Finally, it can be shown that the condition number of the matrix system is $O(h^{-2})$.

**2.2. Norm-equivalence vs. practicality.** Since (2) and (3) imply that

$$\alpha_1 \|u\|_S^2 \leq J(u; 0, 0) \leq \alpha_2 \|u\|_S^2, \tag{14}$$

we refer to the functional $J(\,\cdot\,;\,\cdot,\,\cdot)$ as being *norm equivalent.* This property of the functional causes the LSFEM defined by (5) to recover all the desirable properties of the Rayleigh–Ritz setting. However, the norms that enter the definition of the functional $J(\,\cdot\,;\,\cdot,\,\cdot)$ as well as the form of the PDE system (1) can render the resulting LSFEM impractical. Thus, in order to define a practical LSFEM, one may have to define a least-squares functional that is not norm equivalent in the sense of (14). We take up this issue in §3. Here, we examine the examples of §2.1 to see what guidance they give us about what makes an LSFEM practical.

**2.2.1. First-order system form of the PDEs.** Perhaps the most important observation that can be made from the examples of §2.1 is that the example of §2.1.2 involved a first-order system of PDEs and an LSFEM that allowed for the easy construction of finite element bases (because one could work with merely continuous finite element spaces) and resulted in matrix systems with relative good conditioning. As a result, all modern LSFEMs are based on first-order formulations of PDE systems. Of course, many if not most PDEs of practical interest are not usually posed as first-order systems. Thus, *the first step in defining an LSFEM should be recasting a given PDE system into a first-order system.*

Unfortunately, there is no unique way to do this. For example, the three problems

$$\left\{\begin{array}{ll} \boldsymbol{u} + \nabla\phi = \boldsymbol{0} & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \phi = 0 & \text{on } \Gamma \end{array}\right\}, \quad \left\{\begin{array}{ll} \boldsymbol{u} + \nabla\phi = \boldsymbol{0} & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \nabla \times \boldsymbol{u} = \boldsymbol{0} & \text{in } \Omega \\ \phi = 0 & \text{on } \Gamma \end{array}\right\}, \quad \left\{\begin{array}{ll} \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \nabla \times \boldsymbol{u} = \boldsymbol{0} & \text{in } \Omega \\ \boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{0} & \text{on } \Gamma \end{array}\right\}$$

are all first-order systems that are equivalent to the Poisson problem (11). Each happens to be norm equivalent, but with respect to different norms. If we assume that in each case the boundary condition is imposed on the solutions space, we have that, for the three problems, the space $S$ in (2) is respectively given by $H_0^1(\Omega) \times H(\Omega, \text{div})$, $H_0^1(\Omega) \times \boldsymbol{H}^1(\Omega)$, and $\boldsymbol{H}_\tau^1(\Omega)$, where $H(\Omega, \text{div}) = \{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \mid \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\}$ and $\boldsymbol{H}_\tau^1(\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) \mid \boldsymbol{n} \times \boldsymbol{v} = \boldsymbol{0} \text{ on } \Gamma\}$.

**2.2.2. Functionals formed using $L^2$ norms of equation residuals.** Another observation that can be gleaned from the examples of §2.1 is that if one wants to be able to assemble the matrix system using standard finite element techniques, *then one should use $L^2$ norms of equation residuals in the definition of the least-squares functional.* Unfortunately, it is not always the case that the resulting least-squares functional is norm equivalent. Let us explore this issue in more detail.

Consider the Stokes problem

$$-\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla \cdot \boldsymbol{u} = 0 \ \text{ in } \Omega \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{0} \ \text{ on } \Gamma. \tag{15}$$

The most popular LSFEM for this problem is based on the first-order system

$$\nabla \times \boldsymbol{\omega} + \nabla p = \boldsymbol{f}, \quad \boldsymbol{\omega} = \nabla \times \boldsymbol{u}, \quad \nabla \cdot \boldsymbol{u} = 0 \ \text{ in } \Omega \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{0} \ \text{ on } \Gamma \tag{16}$$

that is known for obvious reasons as the *velocity–vorticity–pressure formulation.* One would then be tempted to use the functional

$$J_0(\boldsymbol{u}, \boldsymbol{\omega}, p; \boldsymbol{f}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_0^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2 \tag{17}$$

that involves only $L^2(\Omega)$ norms of equation residuals. Indeed, this is the most popular approach for defining LSFEM for the Stokes equations. Unfortunately, the functional (17) is not norm equivalent [13]. On the other hand, the functional

$$J_{-1}(\boldsymbol{u}, \boldsymbol{\omega}, p; \boldsymbol{f}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_{-1}^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2$$

is equivalent to $\|\boldsymbol{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2$ [13]. So, on the one hand, the lack of norm equivalence for the functional $J_0(\cdot\,,\ \cdot\,,\ \cdot\,;\ \cdot)$ results in a loss of accuracy of the LSFEM approximations based on that functional. On the other hand, the appearance of the $\boldsymbol{H}^{-1}(\Omega)$ norm in the functional $J_{-1}(\cdot\,,\ \cdot\,,\ \cdot\,;\ \cdot)$ results in an impractical LSFEM because the matrix systems are not easily assembled.[10]

## 3. More sophisticated LSFEMs

To define the least-squares principle (4), one had to choose the pair $\{S,\ J(\,\cdot\,;\ f,\ g)\}$, where $S$ denotes a solution Hilbert space and $J(\,\cdot\,;\ f,\ g)$ a functional defined over $S$ that satisfies the norm-equivalence relation (14). We refer to the variational principle (4) as the *continuous* least-squares principle.[11] The straightforward LSFEM was defined by choosing a finite element subspace $S^h \subset S$ and then minimizing the functional $J(\,\cdot\,;\ f,\ g)$ over $S^h$; see (5). We refer to the straightforward LSFEM as the *conforming* LSFEM. For such LSFEMs, we obtain the error estimate (10).

[10] A similar dilemma arises when one imposes boundary conditions through the least-squares functional.

[11] Here, "continuous" refers to the fact that solutions of (4) are also solutions of the PDE system (1). Recall also that (14) follows from the well-posedness relation (2) for the PDE system.

Conforming LSFEMs can be generalized so that their applicability and practicality are enhanced. Here, we briefly discuss some of these generalizations. We still have in mind approximating solutions of the continuous least-squares principle (4) or what is equivalent, solutions of the PDE system (1). We again choose a finite element space $S^h$ and a convex, quadratic functional $J_h(\cdot\,;f,g)$ defined over $S^h$. The pair $\{S^h, J_h(\cdot\,;f,g)\}$ gives rise to the *discrete* least-squares principle

$$\min_{u^h \in S^h} J_h(u^h; f, g). \tag{18}$$

Since we only require that the functional $J_h(\cdot\,;f,g)$ be defined for functions in $S^h$, we refer to LSFEMs constructed in this manner as *discrete* LSFEMs.

The functional $J_h(\cdot\,;f,g)$ is required to satisfy the following non-restrictive assumptions.

H1. There exists a *discrete energy inner product* $(\cdot,\cdot)_h : S^h \times S^h \mapsto \Re$ and a *discrete energy norm* $\|\cdot\|_h = (\cdot,\cdot)_h^{1/2}$ such that $J_h(u^h;0,0) = (u^h,u^h)_h = \|u^h\|_h^2$ for all $u^h \in S^h$.

H2. There exist bilinear forms $E(\cdot\,,\cdot)$ and $T(\cdot\,,\cdot)$ such that for all smooth functions $u \in S$ and all $u^h \in S^h$

$$J_h(u^h; \mathcal{L}u, \mathcal{R}u) = \|u - u^h\|_h^2 + E(u, u^h) + T(u,u). \tag{19}$$

The two assumptions are sufficient to prove the following results about solutions of (18).

**Theorem 3.1.** *Assume that hypotheses* H1 *and* H2 *hold for the discrete principle* $\{S^h, J_h(\cdot\,;f,g)\}$ *and let* $u$ *denote a sufficiently smooth solution of* (1). *Then the problem* (18) *has a unique solution* $u^h \in S^h$. *Moreover,* $u^h$ *satisfies*

$$\|u - u^h\|_h \le \inf_{v^h \in S^h} \|u - v^h\|_h + \sup_{v^h \in S^h} \frac{E(u, v^h)}{\|v^h\|_h}. \tag{20}$$

A discrete least-squares functional $J_h(\cdot\,;f,g)$ will be referred to being *order r-consistent* if there exists a positive number $r$ such that for all sufficiently smooth functions $u \in S$, the second term on the right-hand side of (20) can be bounded from above by $C(u)h^r$, where $C(u)$ is a positive number whose value may depend on $u$ but not on $h$. If $J_h(\cdot\,;f,g)$ is order $r$-consistent, then (20) implies that

$$\|u - u^h\|_h \le \inf_{v^h \in S^h} \|u - v^h\|_h + C(u)h^r. \tag{21}$$

Theorem 3.1 shows that discrete LSFEMs can work under a minimal set of assumptions. It also explains why LSFEMs tend to be much more robust than their mixed FEM counterparts; unlike the inf-sup conditions that are required for the latter type of method, defining pairs $\{S^h, J_h(\cdot\,;f,g)\}$ such that the assumptions H1 and H2 are satisfied is not a difficult task.

**Constructing discrete least-squares functionals.** Theorem 3.1 provides estimates for the error with respect to the discrete norm $\|\cdot\|_h$. Of greater interest is estimating errors using the (mesh-independent) solution norm $\|\cdot\|_S$ associated with the PDE problem (1). Since $S^h \subset S$, it is certainly true that $\|\cdot\|_S$ acts as another norm on $S^h$, in addition to $\|\cdot\|_h$. Thus, since $S^h$ is finite dimensional, these two norms are comparable. However, the comparability constants may depend on $h$; if they do, then error estimates analogous to (20) and (21) but in terms of the norm $\|\cdot\|_S$ will surely involve constants that depend on inverse powers of $h$ and, at the least, accuracy may be compromised. We conclude that hypotheses H1 and H2 do not sufficiently connect $J_h(\cdot\,; f, g)$ to the problem (1) for us to determine much about the properties of the error in the discrete LSFEM solution with respect to $\|\cdot\|_S$ norm. Thus, we now discuss how to construct discrete least-squares functionals so that we can get a handle on these properties.

We assume that (2) and (14) hold for the problem (1), the least-squares functional $J(\cdot\,; f, g)$, the solution space $S$, and the data spaces $H_\Omega$ and $H_\Gamma$. Let $\mathcal{D}_S$, $\mathcal{D}_\Omega$, and $\mathcal{D}_\Gamma$ denote *norm-generating* operators that allow us to relate the norms on $S$, $H_\Omega$, and $H_\Gamma$, respectively, to[12] $L^2(\Omega)$ norms, i.e., such that, for all $u \in S$, $f \in H_\Omega$, and $g \in H_\Gamma$, $\|u\|_S = \|\mathcal{D}_S u\|_0$, $\|f\|_{H_\Omega} = \|\mathcal{D}_\Omega f\|_0$, and $\|g\|_{H_\Gamma} = \|\mathcal{D}_\Gamma g\|_{0,\Gamma}$. We then let

$$J_h(u^h; f, g) = \|\mathcal{D}_\Omega^h(\mathcal{L}^h u^h - \mathcal{Q}_\Omega^h f)\|_0^2 + \|\mathcal{D}_\Gamma^h(\mathcal{R}^h u^h - \mathcal{Q}_\Gamma^h g)\|_0^2,$$

where $\mathcal{D}_\Omega^h$, $\mathcal{D}_\Gamma^h$, $\mathcal{L}^h$, and $\mathcal{R}^h$ are approximations of the operators $\mathcal{D}_\Omega$, $\mathcal{D}_\Gamma$, $\mathcal{L}$, and $\mathcal{R}$, respectively, and $\mathcal{Q}_\Omega^h : H_\Omega \mapsto L^2(\Omega)$ and $\mathcal{Q}_\Gamma^h : H_\Gamma \mapsto L^2(\Gamma)$ are projections. It can be shown that $J_h(u^h; f, g)$ satisfies (19) with a specific form for $E(u, v^h)$.

The operators $\mathcal{L}$ and $\mathcal{R}$ define the problem (1) that is being solved so that the main objective in choosing $\mathcal{L}^h$ and $\mathcal{R}^h$ is to make $J_h(u; f, g)$ as small as possible for the exact solutions $u$. An appropriate choice is to use operators that will lead to truncation errors of order $r$ in (19), i.e., $\mathcal{L}^h$ and $\mathcal{R}^h$ should be such that (21) holds. On the other hand, $\mathcal{D}_\Omega$ and $\mathcal{D}_\Gamma$ define the energy balance of (1), i.e., the proper scaling between data and solution spaces. As a result, the main objective in the choice of $\mathcal{D}_\Omega^h$ and $\mathcal{D}_\Gamma^h$ is to ensure that the scaling induced by $J_h(\cdot\,; f, g)$ is as close as possible to (2), i.e., to "bind" $\{S^h, J_h(\cdot\,; f, g)\}$ to the energy balance of $\{S, J(\cdot\,; f, g)\}$.

For *norm-equivalent* discrete least-squares principles, $J_h(\cdot, f, g)$ satisfies

$$\hat{\alpha}_1 \|u^h\|_S \le J_h(\cdot\,; 0, 0) \le \hat{\alpha}_2 \|u^h\|_S \quad \text{for all } u^h \in S^h.$$

If the finite element space satisfies standard inverse assumptions, minimizers of this functional satisfy the error estimate

$$\|u - u^h\|_S \le C\left\{ \inf_{v^h \in S^h} \|u - v^h\|_S + \left( \inf_{v^h \in S^h} \|u - v^h\|_h + \sup_{v^h \in S^h} \frac{E(u, v^h)}{\|v^h\|_h} \right) \right\}.$$

[12] Recall from §2.2 that the use of $L^2(\Omega)$ norms in the definition of the least-squares functional is a key factor to making an LSFEM practical.

For *quasi norm-equivalent* discrete least-squares principles, $J_h(\cdot\,;\, f, g)$ satisfies

$$\hat{\alpha}_1^h \|u^h\|_S \leq J_h(\cdot\,; 0, 0) \leq \hat{\alpha}_2^h \|u^h\|_S,$$

where $\hat{\alpha}_1^h > 0$ and $\hat{\alpha}_2^h > 0$ for all $h > 0$ but may depend on $h$. Under additional assumptions, error estimates can also be derived in this case.

## 4. Compatible LSFEMs

Stable mixed finite element methods (MFEMs) for the Poisson equation[13] based on first-order formulations involving a scalar variable $\phi$ and a vector (or flux) variable $\boldsymbol{u}$ require the use of finite element spaces that satisfy an appropriate inf-sup condition [19], [20]. It is well known that pairs of standard, nodal-based, continuous finite element spaces fail the inf-sup condition and lead to unstable mixed methods. It is also well known that the inf-sup condition is circumvented if one uses such simple element pairs in LSFEMs based on $L^2$ least-squares functionals. Ever since such LSFEMs for first-order formulations of the Poisson equation were first considered in [38], this fact has been deemed as an important advantage of those methods over MFEMs. On the other hand, such LSFEMs suffer from two deficiencies. Computationally-based observations indicate that nodal-based LSFEMs do a poor job, compared to stable MFEMs, of conserving mass, i.e., of locally satisfying $\nabla \cdot \boldsymbol{u} = 0$. In addition, excepting in one special case, such methods produce suboptimally accurate (with respect to $L^2(\Omega)$ norms) flux approximations.[14]

Already in [38], optimal $L^2$ error estimates for LSFEMs were established for the scalar variable; however, there and in all subsequent analyses, optimal $L^2$ error estimates for the flux could not be obtained[15] without the addition of a "redundant" curl equation; see, e.g., [23], [24], [26], [39], [43]. Moreover, computational studies in [32] strongly suggested that optimal $L^2$ convergence for flux approximations may in fact be nearly impossible to obtain if one uses pairs of standard, nodal-based, continuous finite element spaces. A notable exception was a case studied in [32] for which optimal $L^2$ error estimates for both the scalar variable and the flux were obtained when these variables were approximated by continuous nodal spaces corresponding to a criss-cross grid. The key to proving these results was the validity of a grid decomposition property (GDP) which was established for the criss-cross grid in [33]. So far, the criss-cross grid remains the only known case of a continuous, nodal-based

[13] Although we consider only the Poisson problem, much of what we discuss can be easily extended to other systems of elliptic PDEs.

[14] The least-squares functionals in question are norm equivalent so that optimally accurate approximations are obtained with respect to the norms for which the equivalences hold. Here, we are interested in error estimates in weaker $L^2(\Omega)$ norms for which the norm equivalence of the least-square functional does not by itself guarantee optimal accuracy.

[15] A somewhat different situation exists for negative-norm-based LSFEMs for which it is known that the $L^2$ accuracy of the flux is optimal with respect to the spaces used; however, for such methods, no error bound for the divergence of the flux could be established; see [18].

finite element space for which the GDP can be verified. More importantly, it was shown in [33] (see also [17]) that the GDP is necessary and sufficient for the stability of MFEMs.

The correlation between the stability of MFEMs and the optimal accuracy of LSFEMs, established in [32], opens up the intriguing possibility that optimal $L^2$ accuracy for the flux may be obtainable for an LSFEM, provided that this variable is approximated using finite element spaces that are stable for an appropriate MFEM. Today, the stability of MFEMs is well understood, and many examples of stable finite element pairs are known. We will show that the use of some of these spaces in an LSFEM indeed can help improve the $L^2$ accuracy of flux approximations.

What we conclude is that if one gives up the use of nodal-based, continuous finite element spaces for the approximation of the flux, one can obtain optimally accurate approximations of the flux with respect to $L^2(\Omega)$ norms. While this conclusion may disappoint the adherents of equal-order implementations,[16] our results do not void LSFEMs as a viable or even preferable computational alternative to MFEMs. To the contrary, they demonstrate that *an LSFEM can be designed that combines the best computational properties of two dual MFEMs* and at the same time manages to avoid the inf-sup conditions and indefinite linear systems that make the latter more difficult to solve. Although we reach this conclusion in the specific context of MFEMs and LSFEMs for the Poisson problem, the idea of defining the latter type of method so that it inherits the best characteristics of a pair of mixed methods that are related through duality may have considerably wider application.

In the rest of this section, we focus the Poisson equation

$$-\Delta\phi = f \ \text{ in } \Omega, \quad \phi = 0 \ \text{ on } \Gamma_d, \quad \text{and} \quad \partial\phi/\partial n = 0 \ \text{ on } \Gamma_n, \tag{22}$$

where $\Omega$ denotes a bounded region in $\Re^d$, $d = 2, 3$, with a Lipschitz continuous boundary $\Gamma$ that consists of two disjoint parts denoted by $\Gamma_d$ and $\Gamma_n$.

**4.1. MFEMs for the Poisson problem.** So as to provide a background for subsequent discussions concerning LSFEMs, we first consider two[17] (dual) MFEMs for the Poisson problem (22) written in the first-order form

$$\nabla \cdot \boldsymbol{u} = f, \quad \boldsymbol{u} + \nabla\phi = \mathbf{0} \ \text{ in } \Omega, \quad \phi = 0 \ \text{ on } \Gamma_d, \quad \boldsymbol{u} \cdot \boldsymbol{n} = 0 \ \text{ on } \Gamma_n. \tag{23}$$

**4.1.1. Stable MFEMs for the Dirichlet principle.** Continuous, nodal finite element spaces built from $m$th degree polynomials, $m \geq 1$, and whose elements satisfy the boundary condition $\phi = 0$ on $\Gamma_n$ are denoted by $\mathcal{S}_m^0$. Note that $\mathcal{S}_m^0 \subset \{\psi \in H^1(\Omega) \mid \psi = 0 \text{ on } \Gamma_d\}$. We denote by $\mathcal{S}_m^1$ the space $\nabla(\mathcal{S}_m^0)$.[18]

[16] Recall that the ability to approximate all variables using simple nodal finite element spaces was one or the advantages of the FEMs in the Rayleigh–Ritz setting that we set out to recover using LSFEMs.

[17] Because they can be derived from two classical optimization problems, we will refer to the two methods as the discretized Dirichlet and Kelvin principles, respectively.

[18] Except for $m = 1$, $\mathcal{S}_m^1$ is not a complete $(m-1)$st degree polynomial space. However, characterizing $\mathcal{S}_m^1$ is not difficult and turns out to be unnecessary in practice.

A stable MFEM based on the Dirichlet principle is defined as follows: seek $\psi_h \in \mathcal{S}_m^0$ and $\boldsymbol{u}_h \in \mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$ such that

$$\begin{cases} \displaystyle\int_\Omega \boldsymbol{u}_h \cdot \boldsymbol{v}_h \, d\Omega + \int_\Omega \nabla\phi_h \cdot \boldsymbol{v}_h \, d\Omega = 0 & \text{for all } \boldsymbol{v}_h \in \mathcal{S}_m^1, \\ \displaystyle\int_\Omega \nabla\psi_h \cdot \boldsymbol{u}_h \, d\Omega = -\int_\Omega f\psi_h \, d\Omega & \text{for all } \psi \in \mathcal{S}_m^0. \end{cases} \tag{24}$$

Note that since $\mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$, even at the discrete level, we may eliminate the flux approximation to obtain the equivalent discrete problem for $\phi_h \in \mathcal{S}_m^0$

$$\int_\Omega \nabla\phi_h \cdot \nabla\psi_h \, d\Omega = \int_\Omega f\psi_h \, d\Omega \quad \text{for all } \psi \in \mathcal{S}_m^0 \tag{25}$$

that we recognize as the standard Galerkin discretization of (22). In fact, (24) and (25) are equivalent in that whenever $\phi_h$ is a solution of (25), then $\phi_h$ and $\boldsymbol{u}_h = \nabla\phi_h$ are a solution pair for (24) and conversely. In this way we see that for (24), i.e., the Dirichlet principle, the required inf-sup condition is completely benign in the sense that it can be avoided by eliminating the flux approximation $\boldsymbol{u}_h$ from (24) and solving (25) instead. The required inf-sup condition is implicitly satisfied by the pair of spaces $\mathcal{S}_m^0$ and $\mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$. If one insists on solving (24), then one needs to explicitly produce a basis for $\mathcal{S}_m^1$; this is easily accomplished.

From either (24) or (25) one obtains, for the Dirichlet principle, that if $\phi \in H^{m+1}(\Omega) \cap H_d^1(\Omega)$, then

$$\|\phi - \phi_h\|_0 \le h^{m+1}\|\phi\|_{m+1} \quad \text{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_h\|_0 = \|\nabla(\phi - \phi_h)\|_0 \le h^m\|\phi\|_{m+1}. \tag{26}$$

**4.1.2. Stable MFEMs for the Kelvin principle.** The $\mathrm{BDM}_k$ and $\mathrm{RT}_k$ spaces on $\Omega$ are built from the individual element spaces defined with respect to a finite element $\mathcal{K}$ in a partition $\mathcal{T}_h$ of $\Omega$

$$\mathrm{BDM}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n \quad \text{and} \quad \mathrm{RT}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n + \boldsymbol{x}\,P_k(\mathcal{K})$$

in a manner that ensures the continuity of the normal component across element boundaries; see [20] for details and definitions of the corresponding element degrees of freedom. Since $\mathrm{BDM}_k$ and $\mathrm{RT}_k$ both contain complete polynomials of degree $k$, their approximation properties in $L^2$ are the same. Since $\mathrm{RT}_k$ also contains the higher-degree polynomial component $\boldsymbol{x}\,P_k(\mathcal{K})$, it approximates the divergence of the flux with better accuracy than does $\mathrm{BDM}_k$. Note, however, that this additional component does not help to improve the $L^2$ accuracy of $\mathrm{RT}_k$ spaces because it does not increase to $k+1$ the order of the complete polynomials contained in $\mathrm{RT}_k$.

In what follows, we will denote by $\mathcal{S}_k^2$ the RT and BDM spaces having *equal approximation orders with respect to the divergence operator*, i.e., we set $\mathcal{S}_k^2 = \{\boldsymbol{v} \in H_n(\Omega, \mathrm{div}) \mid \boldsymbol{v}|_{\mathcal{K}} \in \mathcal{S}_k^2(\mathcal{K})\}$, where $\mathcal{S}_k^2(\mathcal{K})$ is one of the finite element spaces

$\mathrm{RT}_{k-1}(\mathcal{K})$ or $\mathrm{BDM}_k(\mathcal{K})$ and $H_n(\Omega, \mathrm{div}) = \{\boldsymbol{v} \in H_n(\Omega, \mathrm{div}) \mid \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_n\}$. We denote by $\mathcal{S}_k^3$ the space $\nabla \cdot (\mathcal{S}_k^2)$. For characterizations of these spaces, see [20].

A stable MFEM based on the Kelvin principle is defined as follows: we seek $\boldsymbol{u}_h \in \mathcal{S}_k^2$ and $\phi_h \in \mathcal{S}_k^3 = \nabla \cdot (\mathcal{S}_k^2)$ such that

$$\begin{cases} \displaystyle\int_\Omega \boldsymbol{u}_h \cdot \boldsymbol{v}_h \, d\Omega - \int_\Omega \phi_h \nabla \cdot \boldsymbol{v}_h \, d\Omega = 0 & \text{for all } \boldsymbol{v}_h \in \mathcal{S}_k^2, \\ \displaystyle\int_\Omega \psi_h \nabla \cdot \boldsymbol{u}_h \, d\Omega = \int_\Omega f \psi_h \, d\Omega & \text{for all } \psi_h \in \mathcal{S}_k^3. \end{cases} \tag{27}$$

For (27), the required inf-sup condition is much more onerous than for (24) in the sense that defining a pair of stable finite element spaces for the scalar variable and the flux is not so straightforward a matter. We refer to [20] for a proof that $(\mathcal{S}_k^3, \mathcal{S}_k^2)$ is a stable pair for the mixed finite element problem (27). Moreover, one can show [20] that for any sufficiently regular exact solution of (23), one has

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \le C h^r \|\boldsymbol{u}\|_r \quad \begin{cases} \text{for } 1 \le r \le k & \text{if } \mathcal{S}_k^2(\mathcal{K}) = \mathrm{RT}_{k-1}, \\ \text{for } 1 \le r \le k+1 & \text{if } \mathcal{S}_k^2(\mathcal{K}) = \mathrm{BDM}_k, \end{cases} \tag{28}$$

$$\|\nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h)\|_0 \le C h^r \|\nabla \cdot \boldsymbol{u}\|_r \quad \text{for } 1 \le r \le k \tag{29}$$

and

$$\|\phi - \phi_h\|_0 \le C h^r (\|\phi\|_r + \|\boldsymbol{u}\|_r) \quad \text{for } 1 \le r \le k. \tag{30}$$

It is important to note that if one uses continuous, nodal based finite element spaces for both the scalar variable and the flux, then (24) and (27) are identical discrete systems. It is well known that this leads to unstable approximations, so that one cannot use such pairs of finite element spaces in the MFEMs (24) or (27).

**4.1.3. The grid decomposition property.** The following result establishes the GDP for the spaces $\mathcal{S}_k^2$ used for the discretized Kelvin principle (27);[19] for a proof, see [14].

**Theorem 4.1.** *For every $\boldsymbol{u}_h \in \mathcal{S}_k^2$, there exist $\boldsymbol{w}_h, z_h$ in $\mathcal{S}_k^2$ such that*

$$\begin{aligned} &\boldsymbol{u}_h = \boldsymbol{w}_h + z_h, \quad \nabla \cdot z_h = 0, \quad \int_\Omega \boldsymbol{w}_h \cdot z_h \, d\Omega = 0, \quad \textit{and} \\ &\|\boldsymbol{w}_h\|_0 \le C\big(\|\nabla \cdot \boldsymbol{u}_h\|_{-1} + h\|\nabla \cdot \boldsymbol{u}_h\|_0\big). \end{aligned} \tag{31}$$

It was shown in [33] that the GDP, i.e., (31), along with the relation $\mathcal{S}_k^3 = \nabla \cdot (\mathcal{S}_k^2)$, are necessary and sufficient for the stability of the discretized Kelvin principle (27).

[19] An analogous GDP can be defined in the context of the finite element spaces $\mathcal{S}_m^0$ used for the discretized Dirichlet principle (24) but it is trivially satisfied.

**4.2. LSFEMs for the Poisson problem.** An LSFEM for the Poisson problem (22) can be defined based on the quadratic functional

$$J(\phi, \boldsymbol{u};\, f) = \|\nabla \cdot \boldsymbol{u} - f\|_0^2 + \|\nabla\phi + \boldsymbol{u}\|_0^2 \tag{32}$$

and the least-squares principle

$$\min_{(\phi,\boldsymbol{u})\in H_d^1(\Omega)\times H_n(\Omega,\mathrm{div})} J(\phi, \boldsymbol{u};\, f). \tag{33}$$

Note that we have used the first-order form (23) of the Poisson problem and that we use $L^2(\Omega)$ norms to measure the equation residuals. Also, we require the functions in the spaces $H_d^1(\Omega)$ and $H_n(\Omega, \mathrm{div})$ to satisfy the boundary conditions $\phi = 0$ on $\Gamma_d$ and $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ on $\Gamma_n$, respectively. The Euler–Lagrange equations corresponding to (33) are given by: seek $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$ such that

$$B(\{\phi, \boldsymbol{u}\}, \{\psi, \boldsymbol{v}\}) = F(\{\psi, \boldsymbol{v}\}) \quad \text{for all } \{\psi, \boldsymbol{v}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div}), \tag{34}$$

where

$$B(\{\phi, \boldsymbol{u}\}, \{\psi, \boldsymbol{v}\}) = \int_\Omega (\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{v})\, d\Omega + \int_\Omega (\nabla\phi + \boldsymbol{u}) \cdot (\nabla\psi + \boldsymbol{v})\, d\Omega \tag{35}$$

and

$$F(\{\psi, \boldsymbol{v}\}) = \int_\Omega f(\nabla \cdot \boldsymbol{v})\, d\Omega. \tag{36}$$

To define an LSFEM, we restrict (33) to the conforming subspace $\mathcal{S}_m^0 \times \mathcal{S}_k^2 \subset H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$ or, equivalently, restrict (34) to those subspaces to obtain the discrete problem: seek $\{\phi_h, \boldsymbol{u}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ such that

$$B(\{\phi_h, \boldsymbol{u}_h\}, \{\psi_h, \boldsymbol{v}_h\}) = F(\{\psi_h, \boldsymbol{v}_h\}) \quad \text{for all } \{\psi_h, \boldsymbol{v}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2. \tag{37}$$

The next theorem states that the functional (32) is norm equivalent.[20] For a proof, see any of [21], [23], [24], [43].

**Theorem 4.2.** *There exist positive constants $\alpha_1$ and $\alpha_2$ such that for any $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$,*

$$\alpha_1\big(\|\phi\|_1^2 + \|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})}^2\big) \le J(\phi, \boldsymbol{u}; 0) \le \alpha_2\big(\|\phi\|_1^2 + \|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})}^2\big). \tag{38}$$

Thus, the LSFEM defined through (37) is an example of an LSFEM that recovers all the desirable properties of the Rayleigh–Ritz setting, except that by using the finite element spaces $\mathcal{S}_m^0$ and $\mathcal{S}_k^2$, we have forced ourselves to not use continuous, nodal-based finite element spaces for the flux approximation.[21] Because we are using finite element spaces that are compatible for the MFEMs (24) and (27), we refer to the LSFEM defined by (37) as a *compatible* LSFEM.

[20] In the theorem, we have that $\|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})} = (\|\boldsymbol{u}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2)^{1/2}$.

[21] We could, of course, use such spaces for the flux approximation, but, as indicated previously, we would then not be able to obtain optimal error estimates with respect to $L^2(\Omega)$ norms.

**4.2.1. Error estimates in $H^1(\Omega) \times H(\Omega, \mathrm{div})$.** We now review the convergence properties of LSFEMs for the Poisson equation with respect to the $H^1(\Omega) \times H(\Omega, \mathrm{div})$ norm. For a proof of the following theorem, see [14].

**Theorem 4.3.** *Assume that the solution* $\{\phi, \boldsymbol{u}\}$ *of* (34) *satisfies* $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \cap H^{m+1}(\Omega) \times H_n(\Omega, \mathrm{div}) \cap \boldsymbol{H}^{k+1}(\Omega)$ *for some integers* $k, m \geq 1$. *Let* $\{\phi_h, \boldsymbol{u}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ *be the solution of the least-squares finite element problem* (37). *Then there exists a constant* $C > 0$ *such that*

$$\|\phi - \phi_h\|_1 + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{H(\Omega,\mathrm{div})} \leq C\big(h^k \|\boldsymbol{u}\|_{k+1} + h^m \|\phi\|_{m+1}\big). \tag{39}$$

*The error estimate* (39) *remains valid if* $\boldsymbol{u}$ *is approximated in the continuous, nodal-based finite element space* $(P_k(\Omega))^n$.

Theorem 4.3 shows that the errors in $\boldsymbol{u}_h$ and $\phi_h$ are equilibrated when $k = m$ and that $(\mathcal{S}_k^0, \mathcal{S}_k^2)$ has the same asymptotic accuracy in the norm of $H^1(\Omega) \times H(\Omega, \mathrm{div})$ as the $C^0$ pair $(\mathcal{S}_k^0, (P_k)^n)$. For this reason, in the implementation of the LSFEM, one usually chooses the nodal-based pair $(\mathcal{S}_k^0, (P_k)^n)$ because it is easier to implement. Indeed, the ability to use equal-order interpolation has been often cited as a primary reason for choosing to use LSFEMs. Nevertheless, the pair is not flawless because optimal $L^2$ norm errors for the flux approximation have proven impossible to obtain without using the very restrictive criss-cross grid or augmenting (23) with an additional redundant curl constraint equation.[22] Also, as we have already mentioned, numerical studies in [32] indicate that the $L^2$ convergence of the flux is indeed suboptimal with such finite element spaces.

We will see that if the nodal approximation of the flux is replaced by an approximation in $\mathcal{S}_k^2$, it may be possible to recover optimal $L^2$ convergence rates without adding the curl constraint. As in [32], the key to this is the GDP.

**4.2.2. Error estimates in $L^2$.** We assume that the solution of the problem

$$-\Delta\psi = \eta \ \text{ in } \Omega, \quad \psi = 0 \ \text{ on } \Gamma_d, \quad \frac{\partial \psi}{\partial n} = 0 \ \text{ on } \Gamma_n$$

satisfies the regularity estimate $\|\psi\|_{s+2} \leq C\|\eta\|_s$ for $s = 0, 1$ and for all $\eta \in H^s(\Omega)$. This is needed since $L^2$ error estimates are based on duality arguments.

### $L^2$ error estimates for the scalar variable.

**Theorem 4.4.** *Assume that the regularity assumption is satisfied, and assume that the solution* $(\phi, \boldsymbol{u})$ *of* (34) *satisfies* $(\phi, \boldsymbol{u}) \in H_d^1(\Omega) \cap H^{m+1}(\Omega) \times H_n(\Omega, \mathrm{div}) \cap \boldsymbol{H}^{k+1}(\Omega)$

[22] The redundant curl constraint $\nabla \times \boldsymbol{u} = \boldsymbol{0}$, first introduced in the least-squares finite element setting in [26] and subsequently utilized by many others (see, e.g., [21], [23], [24], [39]), renders the least-squares functional norm-equivalent with respect to the $H^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ norm but, in some situations, may unduly restrict the range of the data and should be avoided.

*for some integers $k, m \geq 1$. Let $(\phi_h, \boldsymbol{u}_h) \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ be the solution of the least-squares finite element problem* (37). *Then there exists a constant $C > 0$ such that* $\|\phi - \phi_h\|_0 \leq C(h^{k+1}\|\boldsymbol{u}\|_{k+1} + h^{m+1}\|\phi\|_{m+1})$.

For a proof of this theorem, see [14]. The optimal $L^2$ error bound of Theorem 4.4 for the scalar variable does not require that the finite element space for flux approximations satisfy (31), i.e., the GDP. Thus, it remains valid even when continuous, nodal-based finite element spaces are used for the flux approximations, a result first shown in [38]. On the other hand, we will see that the GDP is needed if one wants to improve the $L^2$ accuracy of the flux.

**$L^2$ error estimate for the flux.** The $L^2$ error estimates for approximations to the flux depend on whether $\mathcal{S}_k^2$ represents the $\mathrm{RT}_{k-1}$ or the $\mathrm{BDM}_k$ family. To this end, we have the following result whose proof may be found in [14].

**Theorem 4.5.** *Assume that the hypotheses of Theorem* 4.4 *hold with $k = m = r$. Then there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq C \begin{cases} h^r(\|\boldsymbol{u}\|_{r+1} + \|\phi\|_{r+1}) & \text{if } \mathcal{S}_r^2(\Omega) = \mathrm{RT}_{r-1}, \\ h^{r+1}(\|\boldsymbol{u}\|_{r+1} + \|\phi\|_{r+1}) & \text{if } \mathcal{S}_r^2(\Omega) = \mathrm{BDM}_r. \end{cases} \tag{40}$$

Consider, for example, the lowest-order case for which $r = 1$, $\mathcal{S}_1^0(\Omega) = P_1$, and $\mathcal{S}_1^2(\Omega)$ is either $\mathrm{RT}_0$ or $\mathrm{BDM}_1$. If the least-squares finite element method is implemented with $\mathrm{RT}_0$ elements, (40) specializes to

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq h(\|\boldsymbol{u}\|_2 + \|\phi\|_2).$$

If instead we use $\mathrm{BDM}_1$ elements, we then obtain the improved error bound

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq h^2(\|\boldsymbol{u}\|_2 + \|\phi\|_2).$$

**4.3. Interpretation of results and mass conservation.** We have seen that an LS-FEM method implemented using equal-order, continuous, nodal-based finite element spaces approximates the scalar variable with the same accuracy (with respect to both $H^1(\Omega)$ and $L^2(\Omega)$ norms) as the Galerkin method (25) (or, equivalently, the mixed method (24) for the Dirichlet principle. However, the approximation properties of the Kelvin principle (27) are only partially inherited in the sense that the accuracy in the approximation to the divergence of the flux is recovered, but the accuracy in the flux approximation itself may be of one order less. This should not be too much of a surprise because continuous, nodal-based finite elements provide stable discretization only for the Dirichlet principle (with the exception of the criss-cross grid; see [32]). While least-squares minimization is stable enough to allow for the approximation of scalar variables and the flux by equal-order, continuous, nodal-based finite element

spaces, it cannot completely recover from the fact that such spaces are unstable for the Kelvin principle.

The key observation from §4.2.2 is that an LSFEM can inherit the best properties of *both* the discretized Dirichlet principle (24) and Kelvin principle (27), provided the scalar variable and the flux are approximated by finite element spaces that are stable with respect to these two principles, respectively. Then least-squares finite element solutions recover the accuracy of the Dirichlet principle for the scalar variable and the accuracy of the Kelvin principle for the flux. In a way, we see that, implemented in this particular manner, the LSFEM represents a balanced mixture of the two principles. In [16], an explanation of this observation using the apparatus of differential form calculus is provided as are the results of several illustrative computational experiments.

Unlike LSFEMs based on the use of continuous, nodal-based finite element spaces for all variables, it can be shown that through a simple local post-processing procedure, the compatible LSFEM inherits the local mass conservation properties of the discretized Kelvin principle (27); see [16] for details.

## 5. Alternative LSFEMs

The LSFEMs considered so far follow variants of the template established in §2: first, spaces $S$, $H_\Omega$, and $H_\Gamma$ that verify (2) are determined, then a least-squares functional (3) is defined by measuring equation residuals in the norms of $H_\Omega$ and $H_\Gamma$ and, finally, an LSFEM is obtained by minimizing (3) over a finite-dimensional subspace $S^h$ of $S$. Here, we provide examples of methods that, while still relying on least-squares notions, deviate in more significant ways from that template.

**5.1. Collocation LSFEMs.** The least-squares optimization steps (3) and (4) precede the discretization step (5). In the broadest sense, *collocation* LSFEM (CLSFEM) are methods [25], [31], [41] that reverse the order of these two steps. They are also known as *point least-squares* or *overdetermined collocation* methods.

Let $\{U_j(\boldsymbol{x})\}_{j=1}^J$ denote a basis for a finite element space. We seek an approximate solution of (1) of the form $u(\boldsymbol{x}) \approx \hat{u}_h(\boldsymbol{x}) = \sum_{j=1}^J c_j U_j(\boldsymbol{x})$, where $\boldsymbol{c} = (c_1, c_2, \ldots, c_J)$ is a vector of unknown coefficients. Then *collocation points* $\{\boldsymbol{x}_i\}_{i=1}^{M_1} \subset \Omega$ and $\{\boldsymbol{x}_i\}_{i=M_1+1}^{M} \subset \Gamma$ are chosen in such a way that the corresponding point residuals $\mathcal{L}\hat{u}_h(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)$ and $\mathcal{R}\hat{u}_h(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)$ are well defined. Then a CLSFEM is defined by minimizing, over $\boldsymbol{c} \in \Re^J$, the discrete functional

$$J_c(\boldsymbol{c}; f, g) = \sum_{i=1}^{M_1} \alpha_i \big(\mathcal{L}\hat{u}_h(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\big)^2 + \sum_{i=M_1+1}^{M} \beta_i \big(\mathcal{R}\hat{u}_h(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\big)^2.$$

The weights $\alpha_i$ and $\beta_i$ can be used to adjust the relative importance of the terms in the functional. The necessary condition for the minimization of $J_c(\,\cdot\,; f, g)$ gives rise

to an $M \times J$ linear system $\mathbb{A}\boldsymbol{c} = \boldsymbol{b}$. If $M = J$, then the method reduces to a standard collocation method. If $M > J$, the solution $\boldsymbol{c}$ is obtained in a least-squares sense by solving the normal equations $\mathbb{A}^T\mathbb{A}\boldsymbol{c} = \mathbb{A}^T\boldsymbol{b}$. If the collocation points and weights correspond to a quadrature rule, then the CLSFEM is equivalent to an LSFEM in which integrals are approximated by a quadrature rule.

Since only a finite set of collocation points belonging to the domain $\overline{\Omega}$ need be specified, collocation LSFEMs are attractive for problems posed on irregularly shaped domains; see [41]. On the other hand, since the normal equations tend to become ill-conditioned, such methods require additional techniques such as scaling or orthonormalization in order to obtain a reliable solution; see [31].

**5.2. Discontinuous LSFEMs.** The LSFEMs of §2, 3, and 4 are defined using a conforming finite element subspace $S^h$ of the solution space $S$. Discontinuous LSFEMs (DLSFEMs) are an alternative approach that use finite element subspaces of $L^2(\Omega)$ that consist of piecewise polynomial functions that are not constrained by inter-element continuity requirements. The degrees of freedom on each element can be chosen independently of each other and the elements can have hanging nodes. These features offer great flexibility in implementing adaptive methods because first, resolution on each element can be adjusted as needed and second, new elements can be added by simple subdivisions of existing elements.

In general, the least-squares problem (4) cannot be restricted to a discontinuous space $S^h$ because it is not a proper subspace of $S$. To take advantage of discontinuous spaces, it is necessary to modify (3) so that it is well defined on the "broken" (with respect to a partition $\mathcal{T}_h$ of the domain $\Omega$) data space $\boldsymbol{S} = \{u \in L^2(\Omega) \mid u \in S(\mathcal{K}) \text{ for all } \mathcal{K} \in \mathcal{T}_h\}$. The first DLSFEMs appeared in [2], [22] as least-squares formulations for interface and transmission problems for the Poisson equation. We follow [22], where a DLSFEM is developed for the problem

$$\begin{cases} \nabla \cdot (a_i \boldsymbol{u}_i) = f_i & \text{and} \quad \boldsymbol{u}_i + \nabla\phi_i = \boldsymbol{0} \text{ in } \Omega_i, \quad i = 1, 2, \\ \phi_i = 0 \text{ on } \Gamma_{i,d} & \text{and} \quad \boldsymbol{u}_i \cdot \boldsymbol{n}_i = 0 \text{ on } \Gamma_{i,n}, \quad i = 1, 2, \\ \phi_1 = \phi_2 & \text{and} \quad a_1\boldsymbol{u}_1 \cdot \boldsymbol{n}_1 + a_2\boldsymbol{u}_2 \cdot \boldsymbol{n}_2 = 0 \quad \text{on } \Gamma_{12} \end{cases} \tag{41}$$

that is a first-order formulation of a transmission problem for the Poisson equation.[23] Here, $\Omega_1$ and $\Omega_2$ are two[24] open subsets of $\Omega$ such that $\overline{\Omega}_1 \cup \overline{\Omega}_2 = \overline{\Omega}$ and $\Omega_1 \cap \Omega_2 = \emptyset$. The set $\Gamma_{12} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ is the *interface* between the two subdomains and $\Gamma_{i,d} = \Gamma_d \cap \overline{\Omega}_i$ and $\Gamma_{i,n} = \Gamma_n \cap \overline{\Omega}_i$, $i = 1, 2$.

In the conforming case, an LSFEM for the Poisson equation was defined by using the functional (32) and conforming subspaces of the solution space $S = H^1_d(\Omega) \times H_n(\Omega, \text{div})$. For the problem (41), we instead use the "broken" (with respect to the

[23]The functions $a_1$ and $a_2$ denote a "media property" that is discontinuous across $\Gamma_{12}$.

[24]The generalization to more than two subdomains is straightforward.

partition $\{\Omega_1,\Omega_2\}$) solution space $\boldsymbol{S} = \boldsymbol{H}^1_d(\Omega) \times \boldsymbol{H}_n(\Omega, \mathrm{div})$, where

$$\begin{aligned}
&\boldsymbol{H}^1_d(\Omega) = \{\widetilde{\phi} = \{\phi_1, \phi_2\} \mid \phi_i \in H^1_d(\Omega_i),\ i = 1, 2\} && \text{for the scalar variable,}\\
&\boldsymbol{H}_n(\Omega, \mathrm{div}) = \{\widetilde{\boldsymbol{u}} = \{\boldsymbol{u}_1, \boldsymbol{u}_2\} \mid \boldsymbol{u}_i \in H_n(\Omega_i, \mathrm{div}),\ i = 1, 2\} && \text{for the flux.}
\end{aligned}$$

To define a DLSFEM, we also need to replace (32) by a least-squares functional that can be minimized over $\boldsymbol{S}$. Of course, we also want a functional whose minimizer is a solution of (41). A functional with the desired properties is given by (see [22])

$$\begin{aligned}
J_{12}(\widetilde{\phi}, \widetilde{\boldsymbol{u}};\, f_1,\, f_2) = \sum_{i=1}^{2} \left( \|\nabla \cdot (a_i \boldsymbol{u}_i) - f_i\|^2_{0,\Omega_i^h} + \|\boldsymbol{u}_i + \nabla \phi_i\|^2_{0,\Omega_i^h} \right) \\
+ |\phi_1 - \phi_2|^2_{1/2,\Gamma_{12}} + \|a_1 \boldsymbol{u}_1 \cdot \boldsymbol{n}_1 + a_2 \boldsymbol{u}_2 \cdot \boldsymbol{n}_2\|^2_{-1/2,\Gamma_{12}}.
\end{aligned} \tag{42}$$

Interface terms in (42) are treated in exactly the same way as one would impose weak Dirichlet and Neumann conditions, respectively. To obtain a practical method, they are replaced by weighted $L^2$ norms on $\Gamma_{12}$. Choosing $\boldsymbol{S}^h \subset \boldsymbol{S}$ completes the formulation of the DLSFEM; see [22] for further details.

The Treffetz element least-squares method [42], [45] can be viewed as a variant of the DLSFEM. The term "Treffetz elements" usually refers to methods that use approximation spaces consisting of piecewise analytic solutions of the PDE. Such spaces provide highly accurate approximations of the broken solution space $\boldsymbol{S}$ so that they also require functionals that are well-posed with respect to that space. Given a Treffetz element space, it is a trivial matter to use (42) to define a DLSFEM; see [42], [45] for further details.

## 6. Open problems in LSFEM

We close with a brief discussion of some of the open problems that exist in the theory and application of LSFEMs.

**6.1. Hyperbolic PDEs.** Recovery of the Rayleigh–Ritz properties by LSFEMs relies on the existence of Hilbert spaces that validate the bounds (2) for (1). Such bounds are natural for elliptic PDEs and can be derived for any such PDE by using the Agmon–Douglis–Nirenberg theory [1]. On the other hand, for *hyperbolic* PDEs such bounds are not so natural, partly because they admit data in $L^p$ spaces and their solutions may have contact discontinuities and shock waves.

Recall that (7) can be viewed as a Galerkin method applied to a higher-order PDE. As a result, LSFEMs for hyperbolic equations designed using a Hilbert space setting are equivalent to a Galerkin discretization of a *degenerate* elliptic PDE. The result is an LSFEM that will have excellent stability properties but which will smear shocks and discontinuities; see [11] for numerical examples.

To illustrate some of the pitfalls that can be encountered with hyperbolic PDEs, it suffices to consider the simple linear convection-reaction problem

$$\nabla \cdot (\boldsymbol{b}u) + cu = f \ \text{ in } \Omega \quad \text{and} \quad u = g \ \text{ on } \Gamma_-, \tag{43}$$

where $\boldsymbol{b}$ is a given convection vector, $c(\boldsymbol{x})$ is a bounded measurable function on $\Omega$, and $\Gamma_- = \{\boldsymbol{x} \in \Gamma \mid \boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{b}(\boldsymbol{x}) < 0\}$ is the inflow part of the boundary $\Gamma$. A straightforward $L^2(\Omega)$ norm-based least-squares principle for (43) is defined by minimizing the functional

$$J(v; f, g) = \|\nabla \cdot (\boldsymbol{b}v) + cv - f\|_0^2 + \|v - g\|_{0,\Gamma_-}^2 \tag{44}$$

over the Hilbert space $S = \{u \in L^2(\Omega) \mid \mathcal{L}u = \nabla \cdot (\boldsymbol{b}v) + cv \in L^2(\Omega)\}$. Then the following theorem can be obtained [10].

**Theorem 6.1.** *Assume that $\Gamma_-$ is non-characteristic and $c + \frac{1}{2}\nabla \cdot \boldsymbol{b} \geq \sigma > 0$ for some constant $\sigma$. Then $J(v; 0, 0) = \|\nabla \cdot (\boldsymbol{b}v) + cv\|_0^2 + \|v\|_{0,\Gamma_-}^2$ is equivalent to the graph norm $\|v\|_S^2 = \|v\|_0^2 + \|\mathcal{L}v\|_0^2$. For every $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_-)$, (44) has a unique minimizer $u \in S$ and for that $u$ we have that $J(u; 0, 0) \leq \|f\|_0 + \|g\|_{0,\Gamma_-}^2$.*

This theorem shows that if the data belongs to $L^2$, all the prerequisites needed to define an LSFEM are fulfilled. We can proceed as in §2 and define a method in the most straightforward way by restricting the Euler–Lagrange equation corresponding to the minimization of (44) to a finite dimensional subspace $S^h \subset S$.

However, the convection-reaction problem (43) is meaningful even if the data[25] $f$ belongs only to the Banach space $L^1(\Omega)$. In this case, proper solution and data spaces for (43) are given by $S = \{v \in L^1(\Omega) \mid \nabla \cdot (\boldsymbol{b}v) \in L^1(\Omega)\}$ and $H = L^1(\Omega)$, respectively. One can show [35] that $\mathcal{L}$ is an isomorphism $S \mapsto H$ and so, instead of (2), we have a similar bound but in Banach spaces: $\alpha_1 \|u\|_S \leq \|\mathcal{L}u\|_H$ for all $u \in S$.

Now, consider the unconstrained minimization problem associated with the spaces $S$ and $H$:

$$\min_{u \in S} J_1(u; f), \quad \text{where } J_1(u; f) = \|\mathcal{L}u - f\|_{L^1(\Omega)} = \int_\Omega |\mathcal{L}u - f| \, d\Omega. \tag{45}$$

For our model equation (43), this is the "correct" minimization problem that, restricted to $S^h \subset S$, will have solutions that do not smear discontinuities. This fact has been recognized independently in [40] and more recently in [35]. In [35], it is also shown that under some reasonable assumptions on $S^h$, the discrete problem

$$\min_{u^h \in S^h} J_1(u^h; f) \tag{46}$$

has at least one global minimizer, no local minimizers, and a solution that satisfies the stability bound $\|u^h\|_S \leq C\|f\|_H$.

[25] We assume now that $g = 0$.

We can view (45) as yet another example of the conflict between practicality and optimality. In this case, however, the practicality issue is much more severe because (45) is not differentiable, we cannot write a first-order optimality condition, and the discrete problem (46) does not give rise to a matrix problem. This is the chief reason that so far there are only two examples [35], [40] of FEMs for (43) based on the $L^1$ optimization problem (45). In [35], the minimizer of (46) is approximated by solving a sequence of regularized $L^1$ optimization problems that are differentiable. The method of [40] uses a sequence of more conventional $L^2$ least-squares approaches, but defined using an adaptively weighted $L^2$ inner product. The weights are used to weaken contributions to the least-squares functional from elements that contain solution discontinuities.

At this point, there is very limited experience with solving hyperbolic PDEs by minimizing functionals over Banach spaces. For problems with non-smooth data, computational experiments with the methods of [35] and [40] show that they are superior to LSFEMs defined through the minimization of (44); most notable is their ability to provide sharp discontinuity profiles without over- and under-shooting. A series of experiments in [35] also points strongly towards a possibility that the numerical solutions actually obey a maximum principle on general unstructured grids and that the $L^1$-based algorithm seems to be able to select viscosity solutions. However, at present, there are no mathematical confirmations of these facts, nor is it known whether such algorithms for hyperbolic conservation laws are able to provide accurate shock positions and speeds.

Despite the promise of $L^1$ optimization techniques, the state of LSFEMs for hyperbolic problems is far from satisfactory. Straightforward $L^2$ norm-based LSFEMs are clearly not the most appropriate as they are based on the "wrong" stability estimate for the problem. $L^1$ norm-based techniques give far better results but are more complex and, in the case of [35], require the solution of nonlinear optimization problems. Thus, the jury is still out on whether or not it is possible to define a simple, robust, and efficient LSFEMs for hyperbolic problems that will be competitive with specially designed, upwind schemes employing flux limiters.

**6.2. Mass conservation.** In §4 it was shown that LSFEMs for the Poisson equation can be implemented in a way that allows them to inherit the best computational properties of MFEMs for the same problem. In particular, it is possible to define an LSFEM for (23) so that the approximation *locally conserves mass*.

Currently, the methods in §4 are the only such example. Achieving local mass conservation in LSFEMs for incompressible, viscous flows remains an important open problem. All existing LSFEMs for incompressible, viscous flows conserve mass only approximately so that $\|\nabla \cdot \boldsymbol{u}^h\|_0 = O(h^r)$, where $r$ is the approximation order of the finite element space. For low-order elements, which are among the most popular and easy to use elements, LSFEMs have experienced severe problems with mass conservation. For LSFEMs based on the velocity–vorticity–pressure system (16), these problems were first identified in [27] where also a solution was proposed that

combines least-squares principles *and* Lagrange multipliers to achieve element-wise mass conservation. Then the resulting *restricted* LSFEM treats the continuity equation $\nabla \cdot \boldsymbol{u} = 0$ as an additional constraint that is enforced on each element by a Lagrange multiplier. The method achieves remarkable local conservation but compromises the main motivation underlying LSFEMs: to recover a Rayleigh–Ritz setting for the PDE. In particular, property 4c does not hold.

An alternative to exact local conservation is an LSFEM with *enhanced* total mass conservation. This can be effected by increasing the importance of the continuity residual by using weights. A weighted LSFEM for (16) using the functional

$$J_W(\boldsymbol{\omega}, p, \boldsymbol{u}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_0^2 + \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \big( W \|\nabla \cdot \boldsymbol{u}\|_{0,\mathcal{K}}^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_{0,\mathcal{K}}^2 \big)$$

was studied in [28] where numerical studies showed that fairly a small weight, e.g., $W = 10$, helps to significantly improve total mass conservation.

Thus, for the Stokes problem, at present there are methods that either recover local mass conservation but forfeit some important advantages of the Rayleigh–Ritz settings or retain all those advantages but can at best provide improved global conservation. It is of interest to explore whether or not the ideas of §4 can be extended to develop compatible LSFEMs for viscous flows that retain all the Rayleigh–Ritz advantages and at the same time locally conserve mass.

**6.3. LSFEMs for nonlinear problems.** Consider the nonlinear version of (1)

$$\mathcal{L}u + \mathcal{G}(u) = f \ \text{ in } \Omega \quad \text{and} \quad \mathcal{R}u = g \ \text{ on } \Gamma, \tag{47}$$

where $\mathcal{G}(u)$ is a nonlinear term. Formally, a least-squares principle for (1) can be easily extended to handle (47) by modifying (4) and (3) to

$$\min_{u \in S} J_{\mathcal{G}}(u; f, g), \quad \text{where } J_{\mathcal{G}}(u; f, g) = \|\mathcal{L}u + \mathcal{G}(u) - f\|_{H_\Omega}^2 + \|\mathcal{R}u - g\|_{H_\Gamma}^2 \tag{48}$$

and then define an LSFEM by restricting (48) to a family $S^h \subset S$. While the extension of LSFEMs to (47) is trivial, its analysis is not and remains one of the open problems in LSFEMs. Compared with the well-developed mathematical theory for linear elliptic problems [2], [13], [18], [21], [23], [24], [26], [32], [38], analyses of LSFEMs for nonlinear problems are mostly confined to the Navier–Stokes equations [7]–[9].

It can be shown that the Euler–Lagrange equation associated with the least-squares principle (48) for the Navier–Stokes equations has the abstract form

$$F(\lambda, U) \equiv U + T \cdot G(\lambda, U) = 0, \tag{49}$$

where $\lambda$ is the Reynolds number, $T$ is a least-squares solution operator for the associated Stokes problem, and $G$ is a nonlinear operator. As a result, the corresponding discrete nonlinear problem has the same abstract form

$$F^h(\lambda, U^h) \equiv U^h + T^h \cdot G(\lambda, U^h) = 0, \tag{50}$$

where $T^h$ is an approximation of $T$. The importance of (50) is signified by the fact that discretization in (50) is introduced solely by means of an approximation to the *linear* operator $T$ in (49). As a result, under some assumptions, one can show that the error in the nonlinear approximation defined by (50) is of the same order as the error in the least-squares solution of the linear Stokes problem.

One of the obstacles in extending this approach to a broader class of nonlinear problems is that after the application of a least-squares principle, the (differentiation) order of the nonlinear term may change.

## References


[1] Agmon, S., Douglis, A., Nirenberg, L., Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II. *Comm. Pure Appl. Math.* **17** (1964), 35–92.

[2] Aziz, A., Kellogg, R., Stephens, A., Least-squares methods for elliptic systems. *Math. Comp.* **44** (1985), 53–70.

[3] Babuška, I., Aziz, K., Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (ed. by K. Aziz and I. Babuška), Academic Press, New York 1972.

[4] Barth, T., Bochev, P., Gunzburger, M., Shadid, J., A Taxonomy of consistently stabilized finite element methods for the Stokes problem. *SIAM J. Sci. Comput.* **25** (2004), 1585–1607.

[5] Becker, R., Braack, M., A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo* **38** (2001), 173–199.

[6] Blasco, J., Codina, R., Stabilized finite element method for the transient Navier-Stokes equations based on a pressure gradient projection. *Comput. Methods Appl. Mech. Engrg.* **182** (2000), 277–300.

[7] Bochev, P., Analysis of least-squares finite element methods for the Navier-Stokes equations. *SIAM J. Numer. Anal.* **34** (1997), 1817–1844.

[8] Bochev, P., Cai, Z., Manteuffel, T., McCormick, S., Analysis of velocity-flux least squares methods for the Navier-Stokes equations, Part-I. *SIAM. J. Numer. Anal.* **35** (1998), 990–1009.

[9] Bochev, P., Manteuffel, T., McCormick, S., Analysis of velocity-flux least-squares methods for the Navier-Stokes equations, Part-II. *SIAM. J. Numer. Anal.* **36** (1999), 1125–1144.

[10] Bochev, P., Choi, J., Improved least-squares error estimates for scalar hyperbolic problems. *Comput. Meth. Appl. Math.* **1** (2001), 115–124.

[11] Bochev, P., Choi, J., A comparative numerical study of least-squares, SUPG and Galerkin methods for convection problems. *Int. J. Comput. Fluid Dyn.* **15** (2001), 127–146.

[12] Bochev, P., Dohrmann, C., Gunzburger, M., Stabilization of low-order mixed finite elements for the Stokes equations. *SIAM J. Numer. Anal.* **44** (1) (2006), 82–101.

[13] Bochev, P., Gunzburger, M., Analysis of least-squares finite element methods for the Stokes equations. *Math. Comp.* **63** (1994), 479–506.

[14] Bochev, P., Gunzburger, M., On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles. *SIAM J. Numer. Anal.* **43** (2005), 340–362.

[15] Bochev, P., Gunzburger, M., An absolutely stable pressure-Poisson stabilized method for the Stokes equations. *SIAM J. Numer. Anal.* **42** (2005), 1189–1207.

[16] Bochev, P., Gunzburger, M., Compatible least-squares finite element methods. *SIAM J. Numer. Anal.*, to appear.

[17] Boffi, D., Brezzi, F., Gastaldi, L., On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Math. Comp.* **69** (2000), 121–140.

[18] Bramble, J., Lazarov, R., Pasciak, J., A least squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.* **66** (1997), 935–955.

[19] Brezzi, F., On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO Model. Math. Anal. Numer.* **21** (1974) 129–151.

[20] Brezzi, F., Fortin, M., *Mixed and Hybrid Finite Element Methods*. Springer Ser. Comput. Math. 15, Springer-Verlag, New York 1991.

[21] Cai, Z., Lazarov, R., Manteuffel, T., McCormick, S., First-order system least squares for second-order partial differential equations: Part I. *SIAM J. Numer. Anal.* **31** (1994), 1785–1799.

[22] Cao, Y., Gunzburger, M., Least-squares finite element approximations to solutions of interface problems. *SIAM J. Numer. Anal.* **35** (1998), 393-405.

[23] Carey, G., Pehlivanov, A., Error estimates for least-squares mixed finite elements. *Math. Model Numer. Anal.* **28** (1994), 499–516.

[24] Chang, C.-L., Finite element approximation for grad-div type systems in the plane. *SIAM J. Numer. Anal.* **29** (1992), 452–461.

[25] Chang, C.-L., Gunzburger, M., A subdomain Galerkin/least squares method for first order elliptic systems in the plane. *SIAM J. Numer. Anal.* **27** (1990), 1197–1211.

[26] Chang, C.-L., Gunzburger, M., A finite element method for first order elliptic systems in three dimensions. *Appl. Math. Comp.* **23** (1987), 171–184.

[27] Chang, C.-L., Nelson, J., Least squares finite element method for the Stokes problem with zero residual of mass conservation. *SIAM J. Numer. Anal.* **34** (1997), 480–489.

[28] Deang, J., Gunzburger, M., Issues related to least-squares finite element methods for the Stokes equations. *SIAM J. Sci. Comput.* **20** (1998), 878–906.

[29] Dohrmann, C., Bochev, P., A stabilized finite element method for the Stokes problem based on polynomial pressure projections. *Internat. J. Numer. Methods Fluids* **46** (2004), 183–201.

[30] Douglas, J., Wang, J., An absolutely stabilized finite element method for the Stokes problem. *Math. Comp.* **52** (1989) 495–508.

[31] Eason, E., A review of least-squares methods for solving partial differential equations. *Internat. J. Numer. Methods Engrg.* **10** (1976), 1021–1046

[32] Fix, G., Gunzburger, M., Nicolaides, R., On finite element methods of the least-squares type. *Comput. Math. Appl.* **5** (1979), 87–98.

[33] Fix, G., Gunzburger, M., Nicolaides, R., On mixed finite element methods for first-order elliptic systems. *Numer. Math.* **37** (1981), 29–48.

[34] Girault, V., Raviart, P.-A., *Finite Element Methods for Navier-Stokes Equations*. Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin 1986.

[35] Guermond, J.-L., A finite element technique for solving first-order PDEs in $L^p$. *SIAM J. Numer. Anal.* **42** (2004), 714–737.

[36] Hughes, T., Franca, L., A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity pressure spaces. *Comput. Methods Appl. Mech. Engrg.* **65** (1987), 85–96.

[37] Hughes, T., Franca, L., Balestra, M., A new finite element formulation for computational fluid dynamics: Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Comput. Methods Appl. Mech. Engrg.* **59** (1986), 85–99.

[38] Jesperson, D., A least-squares decomposition method for solving elliptic equations. *Math. Comp.* **31** (1977), 873–880.

[39] Jiang, B.-N., Povinelli, L., Optimal least-squares finite element methods for elliptic problems. *Comput. Methods Appl. Mech. Engrg.* **102** (1993), 199–212.

[40] Jiang, B.-N., Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method. *J. Comput. Phys.* **105** (1993), 108–121.

[41] Liable, J., Pinder, G., Least-squares collocation solution of differential equations on irregularly shaped domains using orthogonal meshes. *Numer. Methods Partial Differential Equations* **5** (1989), 347–361.

[42] Monk, P., Wang, D.-Q., A least-squares method for the Helmholtz equation. *Comput. Meth. Appl. Mech. Engrg.* **175** (1999), 121–136.

[43] Pehlivanov, A., Carey, G., Lazarov, R., Least-squares mixed finite elements for second-order elliptic problems. *SIAM J. Numer. Anal.* **31** (1994), 1368–1377.

[44] Silvester, D., Optimal low order finite element methods for incompressible flow. *Comput. Meth. Appl. Mech. Engrg.* **111** (1994), 357–368.

[45] Stojek, M., Least-squares Treffetz-type elements for the Helmholtz equation. *Internat. J. Numer. Methods Engrg.* **41** (1998), 831–849.

[46] Strang, G., Fix, G., *An Analysis of the Finite Element Method*. Prentice Hall, Englewood Cliffs, NJ, 1973.

Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110, U.S.A.

E-mail: pbboche@sandia.gov

School of Computational Science, Florida State University, Tallahassee, FL 32306-4120, U.S.A.

E-mail: gunzburg@scs.fsu.edu

# A posteriori error analysis and adaptive methods for partial differential equations

Zhiming Chen*

**Abstract.** The adaptive finite element method based on a posteriori error estimates provides a systematic way to refine or coarsen the meshes according to the local a posteriori error estimator on the elements. One of the remarkable properties of the method is that for appropriately designed adaptive finite element procedures, the meshes and the associated numerical complexity are quasi-optimal in the sense that in two space dimensions, the finite element discretization error is proportional to $N^{-1/2}$ in terms of the energy norm, where $N$ is the number of elements of the underlying mesh. The purpose of this paper is to report some of the recent advances in the a posteriori error analysis and adaptive finite element methods for partial differential equations. Emphases will be paid on an adaptive perfectly matched layer technique for scattering problems and a sharp $L^1$ a posteriori error analysis for nonlinear convection-diffusion problems.

**Mathematics Subject Classification (2000).** Primary 65N15; Secondary 65N30.

**Keywords.** A posteriori error estimates, adaptivity, quasi-optimality.

## 1. Introduction

The aim of the adaptive finite element method (AFEM) for solving partial differential equations is to find the finite element solution and the corresponding mesh with least possible number of elements in terms of discrete errors. The task to find the mesh with the desired property is highly nontrivial because the solution is a priori unknown. The basic idea of the seminal work [3] is to find the desired mesh under the principle of error equidistribution, that is, the discretization errors should be approximately equal on each element. The errors on the elements which are also unknown can, however, be estimated by a posteriori error estimates. Today AFEM based on a posteriori error estimates attracts increasing interests and becomes one of the central themes of scientific computation. The purpose of this paper is to report some of the recent advances in the a posteriori error analysis and AFEM for partial differential equations.

A posteriori error estimates are computable quantities in terms of the discrete solution and data, which provide information for adaptive mesh refinement (and coarsening), error control, and equidistribution of the computational effort. We describe here briefly the basic idea of AFEM using the example of solving the Possion equation

*The author is grateful to the support of China National Basic Research Program under the grant 2005CB321701 and the China NSF under the grant 10025102 and 10428105.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

on a polygonal domain $\Omega$ in $\mathbb{R}^2$

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{1.1}$$

Here the source function $f$ is assumed to be in $L^2(\Omega)$. It is well known that the solution of the problem (1.1) may be singular due to the reentrant corners of the domain in which case the standard finite element methods with uniform meshes are not efficient.

Let $\mathcal{M}_h$ be a regular triangulation of the domain $\Omega$ and $\mathcal{B}_h$ be the collection of all inter-element sides of $\mathcal{M}_h$. Denote by $u_h$ the piecewise linear conforming finite element solution over $\mathcal{M}_h$. For any inter-element side $e \in \mathcal{B}_h$, let $\Omega_e$ be the collection of two elements sharing $e$ and define the local error indicator $\eta_e$ as

$$\eta_e^2 := \sum_{K\in\Omega_e} \| h_K f \|^2_{L^2(K)} + \| h_e^{1/2} J_e \|^2_{L^2(e)},$$

where $h_K := \text{diam}(K)$, $h_e := \text{diam}(e)$, and $J_e := [\![ \nabla u_h ]\!]_e \cdot \nu$ stands for the jump of flux across side $e$ which is independent of the orientation of the unit normal $\nu$ to $e$. The following a posteriori error estimate is well known [2]:

$$\| u - u_h \|^2_{H^1(\Omega)} \le C \sum_{e\in\mathcal{B}_h} \eta_e^2.$$

That $\eta_e$ really indicates the error locally is explained by the following local lower bound [39]:

$$\eta_e^2 \le C \sum_{K\in\Omega_e} \| u - u_h \|^2_{L^2(K)} + C \sum_{K\in\Omega_e} \| h_K (f - f_K) \|^2_{L^2(K)},$$

where $f_K = \frac{1}{|K|}\int_K f dx$.

Based on the local error indicator, the usual adaptive algorithm solving the elliptic problem (1.1) reads as follows:

$$\text{Solve} \;\rightarrow\; \text{Estimate} \;\rightarrow\; \text{Refine}.$$

The important convergence property, which guarantees the iterative loop terminates in finite steps starting from any initial coarse mesh, is proved in [23], [30]. It is also widely observed that for appropriately designed adaptive finite element procedures, the meshes and the associated numerical complexity are quasi-optimal in the sense that

$$\| \nabla(u - u_h) \|_{L^2(\Omega)} \approx C N^{-1/2} \tag{1.2}$$

is valid asymptotically, where $N$ is the number of elements of the underlying finite element mesh. Since the nonlinear approximation theory [5] indicates that $N^{-1/2}$ is the highest attainable convergence order for approximating functions in $H^1(\Omega)$ in

two space dimensions over a mesh with $N$ elements, one concludes that AFEM is an *optimal* discretization method for solving the elliptic problem (1.1).

In Section 2 we consider to use AFEM to solve the Helmholtz-type scattering problems with perfectly conducting boundary

$$\Delta u + k^2 u = 0 \qquad \text{in } \mathbb{R}^2 \backslash \bar{D}, \tag{1.3a}$$

$$\frac{\partial u}{\partial \boldsymbol{n}} = -g \qquad \text{on } \Gamma_D, \tag{1.3b}$$

$$\sqrt{r}\left(\frac{\partial u}{\partial r} - \boldsymbol{i}ku\right) \to 0 \qquad \text{as } r = |x| \to \infty. \tag{1.3c}$$

Here $D \subset \mathbb{R}^2$ is a bounded domain with Lipschitz boundary $\Gamma_D$, $g \in H^{-1/2}(\Gamma_D)$ is determined by the incoming wave, and $\boldsymbol{n}$ is the unit outer normal to $\Gamma_D$. We assume the wave number $k \in \mathbb{R}$ is a constant. We study an *adaptive perfectly matched layer* (APML) technique to deal with the Sommerfeld radiation condition (1.3c) in which the PML parameters such as the thickness of the layer and the fictitious medium property are determined through sharp a posteriori error estimates. The APML technique combined with AFEM provides a complete numerical strategy for solving the scattering problem in the framework of finite element which has the nice property that the total computational costs are insensitive to the thickness of the PML absorbing layers. The quasi-optimality of underlying FEM meshes is also observed.

Things become much more complicated when applying AFEM to solve time-dependent partial differential equations. One important question is if one should use the *adaptive method of lines* (AML) in which variable timestep sizes (but constant at each time step) and variable space meshes at different time steps are assumed, or one should consider the *space-time adaptive method* in which space-time domain is considered as a whole and AFEM is used without distinguishing the difference of time and space variables. Our recent studies in [9], [10], [11] reveal that with sharp a posteriori error analysis and carefully designed adaptive algorithms, the AML method produces the very desirable quasi-optimal decay of the error with respect to the computational complexity

$$|\!|\!| u - U |\!|\!|_{\Omega\times(0,T)} \leq C M^{-1/3} \tag{1.4}$$

for a large class of convection-diffusion parabolic problems in two space dimensions using backward Euler scheme in time and conforming piecewise linear finite elements in space. Here $|\!|\!| u - U |\!|\!|_{\Omega\times(0,T)}$ is the energy norm of the error between the exact solution $u$ and the discrete solution $U$, and $M$ is the sum of the number of elements of the space meshes over all time steps. Thus if one takes the quasi-optimality of the computational complexity as the criterion to assess the adaptive methods, then the space-time adaptive method which is less studied in the literature will not have much advantage over the AML method.

A posteriori error analysis for parabolic problems in the framework of AML has been studied intensively in the literature. The main tool in deriving a posteriori error

estimates in [25], [26], [14], [31], [7] is the analysis of linear dual problems of the corresponding *error* equations. The derived a posteriori error estimates, however, depend on the $H^2$ regularity assumption on the underlying elliptic operator. Without using this regularity assumption, energy method is used in [34], [9] to derive an a posteriori error estimate for the total energy error of the approximate solution for linear heat equations. A lower bound for the local error is also derived for the associated a posteriori error indicator in [34], [9]. In [9] an adaptive algorithm is constructed which at each time step, is able to reduce the error indicators (and thus the error) below any given tolerance within finite number of iteration steps. Moreover, the adaptive algorithm is quasi-optimal in terms of energy norm. In [10] an quasi-optimal AML method in terms of the energy norm is constructed for the linear convection-dominated diffusion problems based on $L^1$ a posteriori error estimates.

In Section 3 we study the AML method for the initial boundary value problems of nonlinear convection-diffusion equations of the form

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) - \Delta A(u) = g.$$

We derive sharp $L^\infty(L^1)$ a posteriori error estimates under the non-degeneracy assumption $A'(s) > 0$ for any $s \in \mathbb{R}$. The problem displays both parabolic and hyperbolic behavior in a way that depends on the solution itself. It is discretized implicitly in time via the method of characteristic and in space via continuous piecewise linear finite elements. The analysis is based on the Kružkov "doubling of variables" device and the recently introduced "boundary layer sequence" technique to derive the entropy error inequality on bounded domains. The derived a posteriori error estimate leads to a quasi-optimal adaptive method in terms of the $L^\infty(L^1)$ norm of the error.

## 2. The APML technique for scattering problems

In this section we consider the APML technique for the scattering problem (1.3a)–(1.3c). Since [4] proposed a PML technique for solving the time dependent Maxwell equations, various constructions of PML absorbing layers have been proposed and studied in the literature [38], [37]. Here we introduce the PML technique for (1.3a)–(1.3c) following the method in [19].

Let $D$ be contained in the interior of the circle $B_R = \{x \in \mathbb{R}^2 : |x| < R\}$. In the domain $\mathbb{R}^2 \backslash \bar{B}_R$, the solution $u$ of (1.3a)–(1.3c) can be written under the polar coordinates as follows:

$$u(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{H_n^{(1)}(kr)}{H_n^{(1)}(kR)} \hat{u}_n e^{in\theta}, \quad \hat{u}_n = \frac{1}{2\pi}\int_0^{2\pi} u(R,\theta) e^{-in\theta}\, d\theta. \tag{2.1}$$

where $H_n^{(1)}$ is the Hankel function of the first kind and order $n$. The series in (2.1) converges uniformly for $r > R$ [20].

The basic idea of PML technique is to surround the fixed domain $\Omega_R = B_R \backslash \bar{D}$ with a PML layer of thickness $\rho - R$ and choose the fictitious medium property so that either the wave never reaches its external boundary or the amplitude of the reflected wave is so small that it does not essentially contaminate the solution in $\Omega_R$.

Let $\alpha = 1 + \boldsymbol{i}\sigma$ be the model medium property satisfying $\sigma \in C(\mathbb{R})$, $\sigma \geq 0$, and $\sigma = 0$ for $r \leq R$. The most widely used model medium property $\sigma$ in the literature is the power function, that is,

$$\sigma = \sigma_0 \left( \frac{r - R}{\rho - R} \right)^m, \quad m \geq 1,\ \sigma_0 > 0 \text{ constant}. \tag{2.2}$$

Denote by $\tilde{r}$ the complex radius defined by

$$\tilde{r} = \tilde{r}(r) = \begin{cases} r & \text{if } r \leq R, \\ \int_0^r \alpha(t)dt = r\beta(r) & \text{if } r \geq R. \end{cases}$$

Since $H_n^{(1)}(z) \sim \sqrt{\frac{2}{\pi z}} e^{\boldsymbol{i}(z - \frac{\pi}{2}n - \frac{\pi}{4})}$ as $|z| \to \infty$, [19] obtained the PML equation by considering the following extension of $u$ in the exterior domain $\mathbb{R}^2 \backslash \bar{B}_R$:

$$w(r, \theta) = \sum_{n \in \mathbb{Z}} \frac{H_n^{(1)}(k\tilde{r})}{H_n^{(1)}(kR)} \hat{u}_n e^{\boldsymbol{i}n\theta}, \quad \hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(R, \theta) e^{-\boldsymbol{i}n\theta}\, d\theta. \tag{2.3}$$

It is easy to check that $w$ satisfies

$$\nabla \cdot (A \nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \mathbb{R}^2 \backslash \bar{B}_R,$$

where $A = A(x)$ is a matrix which satisfies, in polar coordinates,

$$\nabla \cdot (A\nabla) = \frac{1}{r} \frac{\partial}{\partial r} \left( \frac{\beta r}{\alpha} \frac{\partial}{\partial r} \right) + \frac{\alpha}{\beta} \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2}.$$

The PML problem then becomes

$$\nabla \cdot (A \nabla \hat{u}) + \alpha\beta k^2 \hat{u} = 0 \quad \text{in } B_\rho \backslash \bar{D}, \tag{2.4a}$$

$$\frac{\partial \hat{u}}{\partial \boldsymbol{n}} = -g \quad \text{on } \Gamma_D,\ \hat{u} = 0 \text{ on } \Gamma_\rho. \tag{2.4b}$$

It is proved in [22], [21] that the resultant PML solution converges exponentially to the solution of the original scattering problem as the thickness of the PML layer tends to infinity. We remark that in practical applications involving PML techniques, one cannot afford to use a very thick PML layer if uniform finite element meshes are used because it requires excessive grid points and hence more computer time and more storage. On the other hand, a thin PML layer requires a rapid variation of the artificial material property which deteriorates the accuracy if too coarse mesh is used in the PML layer.

The APML technique was first proposed in [16] for solving scattering by periodic structures (the grating problem) which uses a posteriori error estimates to determine the PML parameters such as the thickness and the medium property $\sigma_0$ in the (2.2). For the scattering problem (1.3a)–(1.3c), the main difficulty of the analysis is that in contrast to the grating problems in which there are only finite number of outgoing modes, now there are infinite number of outgoing modes expressed in terms of Hankel functions. We overcome this difficulty by the by exploiting the following uniform estimate for the Hankel functions $H_\nu^1$, $\nu \in \mathbb{R}$.

**Lemma 2.1.** *For any* $\nu \in \mathbb{R}$, $z \in \mathbb{C}_{++} = \{z \in \mathbb{C} : \Im(z) \geq 0, \Re(z) \geq 0\}$, *and* $\Theta \in \mathbb{R}$ *such that* $0 < \Theta \leq |z|$, *we have*

$$|H_\nu^{(1)}(z)| \leq e^{-\Im(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}} |H_\nu^{(1)}(\Theta)|.$$

The proof of the lemma which depends on the Macdonald formula for the modified Bessel functions can be found in [12]. Lemma 2.1 allows us to prove the exponentially decaying property of the PML solution without resorting to the integral equation technique in [22] or the representation formula in [21]. As a corollary of Lemma 2.1, we know that the function $w$ in (2.3) satisfies

$$\| w \|_{H^{1/2}(\Gamma_\rho)} \leq e^{-k\Im(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| u \|_{H^{1/2}(\Gamma_R)}.$$

We remark that in [22], [21], it is required that the fictitious absorbing coefficient must be linear after certain distance away from the boundary where the PML layer is placed. We also remark that since (2.5) is valid for all real order $\nu$, the results of [12] can be extended directly to study three dimensional Helmholtz-type scattering problems.

Let $\mathcal{M}_h$ be a regular triangulation of $B_\rho \backslash \bar{D}$ and $u_h$ be the finite element solution of the PML problem (2.4a)–(2.4b). Let $\mathcal{B}_h$ denote the set of all sides that do not lie on $\Gamma_D$ and $\Gamma_\rho^h$. For any $K \in \mathcal{M}_h$, we introduce the residual

$$R_h := \nabla \cdot (A \nabla u_h|_K) + \alpha\beta k^2 u_h|_K.$$

For any interior side $e \in \mathcal{B}_h$ which is the common side of $K_1$ and $K_2 \in \mathcal{M}_h$, we define the jump residual across $e$:

$$J_e := (A \nabla u_h|_{K_1} - A \nabla u_h|_{K_2}) \cdot \nu_e,$$

using the convention that the unit normal vector $\nu_e$ to $e$ points from $K_2$ to $K_1$. If $e = \Gamma_D \cap \partial K$ for some element $K \in \mathcal{M}_h$, then we define the jump residual

$$J_e := 2(\nabla u_h|_K \cdot \boldsymbol{n} + g)$$

For any $K \in \mathcal{M}_h$, denote by $\eta_K$ the local error estimator which is defined by

$$\eta_K = \max_{x \in \tilde{K}} \omega(x) \cdot \left( \|h_K R_h\|_{L^2(K)}^2 + \frac{1}{2} \sum_{e \subset \partial K} h_e \| J_e \|_{L^2(e)}^2 \right)^{1/2},$$

where $\tilde{K}$ is the union of all elements having nonempty intersection with $K$, and

$$\omega(x) = \begin{cases} 1 & \text{if } x \in \bar{B}_R \backslash \bar{D}, \\ |\alpha_0 \alpha| e^{-k\Im(\tilde{r})\left(1-\frac{r^2}{|\tilde{r}|^2}\right)^{1/2}} & \text{if } x \in \bar{B}_\rho \backslash B_R. \end{cases}$$

**Theorem 2.2.** *There exists a constant $C$ depending only on the minimum angle of the mesh $\mathcal{M}_h$ such that the following a posterior error estimate is valid:*

$$\| u - u_h \|_{H^1(\Omega_R)} \le C\Lambda(kR)^{1/2}(1+kR)\Big(\sum_{K\in\mathcal{M}_h} \eta_K^2\Big)^{1/2} + C(1+kR)^2|\alpha_0|^2 e^{-k\Im(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| u_h \|_{H^{1/2}(\Gamma_R)}.$$

*Here* $\Lambda(kR) = \max\left(1, \frac{|H_0^{(1)'}(kR)|}{|H_0^{(1)}(kR)|}\right)$.

From Theorem 2.2 we know that the a posteriori error estimate consists of two parts: the PML error and the finite element discretization error. An adaptive algorithm is developed in [12] which uses the a posteriori error estimate to determine the PML parameters. We first choose $\rho$ and $\sigma_0$ such that the exponentially decaying factor

$$\hat{\omega} = e^{-k\Im(\tilde{\rho})(1-\frac{R^2}{|\tilde{\rho}|^2})^{1/2}} \le 10^{-8},$$

which makes the PML error negligible compared with the finite element discretization errors. Once the PML region and the medium property are fixed, we use the standard finite element adaptive strategy to modify the mesh according to the a posteriori error estimate. The extensive numerical experiments reported in [12] show the competitive behavior of the proposed adaptive method. In particular, the quasi-optimality of meshes is observed and the adaptive algorithm is robust with respect to the choice of the thickness of PML layer: the far fields of the scattering solutions are insensitive to the choices of the PML parameters.

## 3. The AML method for nonlinear convection diffusion problems

Let $\Omega$ is a bounded domain in $\mathbb{R}^d$ $(d = 1, 2, 3)$ with Lipschitz boundary and $T > 0$. In this section we consider the following nonlinear convection-diffusion equation:

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) - \Delta A(u) = g \quad \text{in } Q \tag{3.1}$$

with the initial and boundary conditions

$$u|_{t=0} = u_0, \quad u|_{\partial\Omega\times(0,T)} = 0. \tag{3.2}$$

Here $u = u(x,t) \in \mathbb{R}$, with $(x,t) \in Q = \Omega \times (0,T)$. We assume that the function $f : \mathbb{R} \to \mathbb{R}^d$ is locally Lipschitz continuous, the function $A : \mathbb{R} \to \mathbb{R}$ is nondecreasing and locally Lipschitz continuous, $g \in L^\infty(Q)$ and $u_0 \in L^\infty(\Omega)$.

Problems of the type (3.1) model a wide variety of physical phenomena including porous media flow, flow of glaciers and sedimentation processes, or flow transport through unsaturated porous media which is governed by the so-called Richards equation. For the Richards equation, the existence of weak solutions is considered in [1] and the uniqueness of weak solutions is proved in [33] based on the Kružkov "doubling of variables" technique. Entropy solutions for (3.1) are studied in [6], [29].

The discretization of (3.1) is based on combining continuous piecewise linear finite elements in space with the characteristic finite difference in time. The method of characteristic originally proposed in [24], [35] is widely used to solve convection-diffusion problems in finite element community (cf. e.g. [26], [14]). Given $U_h^{n-1}$ as the finite element approximation of the solution at time $t^{n-1}$, let $\tau_n$ and $V_0^n \subset H_0^1(\Omega)$ be the time step and the conforming linear finite element space at the $n$th time step, then our discrete scheme reads as following: find $U_h^n \in V_0^n$ such that

$$\left\langle \frac{U_h^n - \bar{U}_h^{n-1}}{\tau_n}, v \right\rangle + \langle \nabla A(U_h^n), \nabla v \rangle = \langle \bar{g}^n, v \rangle \quad \text{for all } v \in V_0^n, \tag{3.3}$$

where $\bar{g}^n = \tau_n^{-1} \int_{t^{n-1}}^{t^n} g(x,t)dt$, $\bar{U}_h^{n-1}(x) = U_h^{n-1}(\tilde{X}(t^{n-1}))$, and the approximate characteristic $\tilde{X}(t)$ is defined by

$$d\tilde{X}/dt = f'(U_h^{n-1}(\tilde{X}(t))), \quad \tilde{X}(t^n) = x.$$

The well-known Kružkov "doubling of variables" technique originally appeared in [28] plays a decisive role in the error estimation (both a posteriori and a priori) for numerical schemes solving the Cauchy problems of nonlinear conservation laws (see e.g. [17], [18], [27] and the reference therein). It is also used recently in [32] for the implicit vortex centered finite volume discretization of the Cauchy problems of (3.1) for general non-negative $A'(s) \geq 0$ for all $s \in \mathbb{R}$. The common feature of these studies is that the derived error indicators are of the order $\sqrt{h}$ in the region where the solution is smooth, where $h$ is the local mesh size. We remark that in the region where the diffusion is dominant, the error indicators developed for the parabolic equations (cf. e.g. [34], [9]) are of order $h$. Thus the degeneration of the order of the error indicators used in [32] may cause over-refinements for the solution of (3.1) in the region where the diffusion is dominant.

The basic assumption in this paper is that the diffusion is positive

$$A'(s) > 0, \quad \text{for all } s \in \mathbb{R}.$$

This assumption includes the Richards equation and the viscosity regularization of degenerate parabolic equations, for example, the regularized continuous casting problem which is considered in [14]. The novelty of our analysis with respect to the analysis

for nonlinear conservation laws in [17], [18], [27] or nonlinear degenerate parabolic equations in [32] lies in the following aspects. Firstly, only Cauchy problems are considered in [17], [18], [27], [32]. The difficulty to include boundary condition is essential. Here we use the recently introduced technique of "boundary layer sequence" in [29] to overcome the difficulty. The technique of "boundary layer sequence" allows us to truncate the standard Kružkov test function (see Definition 3.4 below) to obtain the admissible test function in the entropy error identity. Secondly, the nature of the estimators are different: our estimators emphasize the diffusion effect of the problem which requires the assumption $A'(s) > 0$ for any $s \in \mathbb{R}$; the estimates in [32] are valid for any nonlinear function $A$ such that $A'(s) \geq 0$. The nice consequence of the analysis is that our a posteriori error estimates are able to recover the standard sharp a posteriori error estimators in the literature derived for parabolic problem with diffusion coefficients bounded uniformly away from zero.

Now we elaborate the main steps to derive sharp $L^1$ a posteriori error estimate for the discrete scheme (3.3) based on the Kružkov "doubling of variables" device. By testing (3.1) with any function $\varphi \in L^2(0, T; H_0^1(\Omega))$ such that $\phi(\cdot, 0) = \phi(\cdot, T) = 0$, we have

$$\int_0^T \langle \partial_t u, \varphi \rangle \, dt + \int_Q (-f(u) + \nabla A(u)) \cdot \nabla \varphi \, dx dt = \int_Q g\varphi \, dx dt. \tag{3.4}$$

For any $\varepsilon > 0$, let

$$H_\varepsilon(z) = \operatorname{sgn}(z) \min(1, |z|/\varepsilon)$$

be the regularization of the sign function $\operatorname{sgn}(z)$. For any $k \in \mathbb{R}$, define the entropy pair $(U_\varepsilon, F_\varepsilon)$ by

$$U_\varepsilon(z, k) = \int_k^z H_\varepsilon(A(r) - A(k)) \, dr, \quad F_\varepsilon(z, k) = \int_k^z H_\varepsilon(A(r) - A(k)) f'(r) \, dr.$$

The following result is well known (cf. e.g. [6], [29]) by taking $\varphi = H_\varepsilon(A(u) - A(k))\phi$ in (3.4).

**Lemma 3.1.** *For any $\phi \in L^2(0, T; H_0^1(\Omega))$ such that $\phi(\cdot, 0) = \phi(\cdot, T) = 0$, and any $k \in \mathbb{R}$, we have*

$$\begin{aligned} &- \int_Q U_\varepsilon(u, k) \partial_t \phi - \int_Q F_\varepsilon(u, k) \cdot \nabla \phi + \int_Q H_\varepsilon(A(u) - A(k)) \nabla A(u) \cdot \nabla \phi \\ &\qquad + \int_Q H'_\varepsilon(A(u) - A(k)) |\nabla A(u)|^2 \phi \\ &= \int_Q g H_\varepsilon(A(u) - A(k)) \phi. \end{aligned} \tag{3.5}$$

Let $(H^1(\Omega))'$ be the dual space of $H^1(\Omega)$, we define the discrete residual $\mathcal{R} \in L^2(0, T; (H^1(\Omega))')$ through the following relation, for any $\varphi \in H^1(\Omega)$,

$$\langle \partial_t U_h, \varphi \rangle - \langle f(U_h), \nabla \varphi \rangle + \langle \nabla A(U_h), \nabla \varphi \rangle = \langle g, \varphi \rangle - \langle \mathcal{R}, \varphi \rangle. \tag{3.6}$$

For any $k' \in \mathbb{R}$, by taking $\varphi = H_\varepsilon(A(U_h) - A(k'))\phi$ in (3.6), we have the following result.

**Lemma 3.2.** *For any $\phi \in L^2(0, T; H_0^1(\Omega))$ such that $\phi(\cdot, 0) = \phi(\cdot, T) = 0$, and any $k' \in \mathbb{R}$, we have*

$$\begin{aligned} &-\int_Q U_\varepsilon(U_h, k')\partial_t\phi - \int_Q F_\varepsilon(U_h, k')\cdot\nabla\phi \\ &\quad + \int_Q H_\varepsilon(A(U_h) - A(k'))\nabla A(u)\cdot\nabla\phi \\ &\quad + \int_Q H'_\varepsilon(A(U_h) - A(k'))|\nabla A(U_h)|^2\phi \\ &= \int_Q g H_\varepsilon(A(U_h) - A(k'))\phi - \int_0^T \langle \mathcal{R}, H_\varepsilon(A(U_h) - A(k'))\phi\rangle. \end{aligned} \tag{3.7}$$

Now we are going to apply the Kružkov "doubling of variables" technique and will always write $u = u(y, s)$, $U_h = U_h(x, t)$, unless otherwise stated. By taking $k = U_h(x, t)$ in (3.5) and $k' = u(y, s)$ in (3.7), we have the following entropy error identity.

**Lemma 3.3.** *Let $\phi = \phi(x, t; y, s)$ be non-negative function such that*

$$\begin{aligned} &(x, t) \mapsto \phi(x, t; y, s) \in C_c^\infty(Q) \quad \textit{for every } (y, s) \in Q, \\ &(y, s) \mapsto \phi(x, t; y, s) \in C_c^\infty(Q) \quad \textit{for every } (x, t) \in Q. \end{aligned}$$

*Then we have*

$$\begin{aligned} &-\int_{Q\times Q} U_\varepsilon(u, U_h)(\partial_t\phi + \partial_s\phi) - \int_{Q\times Q} F_\varepsilon(u, U_h)(\nabla_x\phi + \nabla_y\phi) \\ &\quad + \int_{Q\times Q} H_\varepsilon(A(u) - A(U_h))\nabla_y A(u)\cdot(\nabla_x\phi + \nabla_y\phi) \\ &\quad + \int_{Q\times Q} H_\varepsilon(A(U_h) - A(u))\nabla_x A(U_h)\cdot(\nabla_x\phi + \nabla_y\phi) \\ &\quad + \int_{Q\times Q} H'_\varepsilon(A(u) - A(U_h))|\nabla_x A(U_h) - \nabla_y A(u)|^2\phi \\ &= -\int_{Q\times Q} \partial_t[U_\varepsilon(U_h, u) - U_\varepsilon(u, U_h))]\phi \\ &\quad - \int_{Q\times Q} \nabla_x[F_\varepsilon(U_h, u) - F_\varepsilon(u, U_h))]\phi \\ &\quad - \int_{Q_{(y,s)}} \int_0^T \langle \mathcal{R}, H_\varepsilon(A(U_h) - A(u))\phi\rangle\, dt. \end{aligned} \tag{3.8}$$

The next objective is to remove the restriction that the test functions in the entropy error identity (3.8) must have vanishing trace. This is achieved by using the technique of boundary layer sequence introduced in [29]. For any $\delta > 0$, the boundary layer sequence $\zeta_\delta$ is defined as the solution of the elliptic problem

$$-\delta^2 \Delta \zeta_\delta + \zeta_\delta = 1 \text{ in } \Omega, \quad \zeta_\delta = 0 \text{ on } \partial\Omega.$$

We specify now the choice of the test function $\phi$ in the entropy error identity (3.8), which is similar to that used in [29].

**Definition 3.4.** Let

$$\phi(x, t, y, s) = \zeta_\delta(x)\zeta_\eta(y)\xi(x, t, y, s)\theta(t),$$

where $\theta \in C_c^\infty(0, T)$ such that $\theta \geq 0$, and $\xi$ is defined as follows. Let $\{\varphi_j\}_{0 \leq j \leq J}$ be a partition of unity subordinate to open sets $B_0, B_1, \ldots, B_J$ such that $\bar{\Omega} \subset \cup_{j=0}^J B_j$, $B_0 \subset\subset \Omega$ and $\partial\Omega \subset \cup_{j=1}^J B_j$. Let $\hat{\varphi}_j \in C_c^\infty(\mathbb{R}^d)$, $0 \leq \hat{\varphi}_j \leq 1$, such that $\text{supp}(\hat{\varphi}_j) \subset B_j$ and $\hat{\varphi}_j(x) = 1$ on the support of $\varphi_j$ so that $\varphi_j(x)\hat{\varphi}_j(x) = \varphi_j(x)$. We use $\varphi_j$ as a function of $y$ and $\hat{\varphi}_j$ as a function of $x$, and denote $\hat{\varphi}_j(x)\varphi_j(y) = \psi_j(x, y)$. Define

$$\xi(x, t, y, s) = \sum_{j=0}^{J} \omega_l(t - s)\omega_m(x' - y')\omega_n(x_d - y_d)\psi_j(x, y),$$

where $\omega_l, \omega_n$ are sequences of symmetric mollifiers in $\mathbb{R}$, $\omega_m$ is a sequence of symmetric mollifier in $\mathbb{R}^{d-1}$, and for $j = 1, 2, \ldots, J$, $x = (x', x_d)$, $y = (y', y_d)$ are local coordinates induced by $\psi_j(x, y)$ in $B_j$, that is, $B_j \cap \partial\Omega = \{x \in B_j : x_d = \rho_j(x')\}$, $B \cap \Omega = \{x \in B_j : x_d < \rho_j(x')\}$ for some Lipschitz continuous function $\rho_j : \mathbb{R}^{d-1} \to \mathbb{R}$.

By taking limit $\delta, \eta \to 0$ in the entropy error identity (3.8), we obtain the following entropy error inequality.

**Theorem 3.5.** *Let $\theta$ and $\xi$ be defined in Definition* 3.4. *Then we have the following entropy error inequality:*

$$\begin{aligned}
&-\int_{Q\times Q} U_\varepsilon(u, U_h)\xi\theta_t - \int_{Q\times Q} K_\varepsilon(u, U_h)\cdot(\nabla_x\xi + \nabla_y\xi)\theta \\
&\quad + \int_{Q\times Q} H_\varepsilon'(A(u) - A(U_h))|\nabla_x A(U_h) - \nabla_y A(u)|^2\xi\theta \\
&\leq -\int_{Q\times Q} \partial_t[U_\varepsilon(U_h, u) - U_\varepsilon(u, U_h))]\xi\theta \\
&\quad - \int_{Q\times Q} \nabla_x[F_\varepsilon(u, U_h) - F_\varepsilon(u, U_h))]\xi\theta
\end{aligned} \tag{3.9}$$

$$
\begin{aligned}
&- \int_{Q_{(y,s)}} \int_{\Sigma_{(x,t)}} \Big( F_\varepsilon(u, U_h) - H_\varepsilon(A(u) - A(U_h)) \nabla_y A(u) \Big) \cdot \nu_x \xi \theta \\
&\quad - \int_{Q_{(x,t)}} \int_{\Sigma_{(y,s)}} \Big( F_\varepsilon(u, U_h) - H_\varepsilon(A(U_h) - A(u)) \nabla_x A(U_h) \Big) \cdot \nu_y \xi \theta \\
&\quad\quad - \int_{Q_{(y,s)}} \int_0^T \langle \mathcal{R}, H_\varepsilon(A(U_h) - A(u)) \xi \theta \rangle \, dt,
\end{aligned}
$$

*where* $K_\varepsilon(u, U_h) = F_\varepsilon(u, U_h) - H_\varepsilon(A(u) - A(U_h))(\nabla_y A(u) - \nabla_x A(U_h))$, $\Sigma = \partial\Omega \times (0, T)$, *and* $\Sigma_{(x,t)}$ *or* $\Sigma_{(y,s)}$ *are the domain of integration of* $\Sigma$ *with respect to* $(x, t)$ *or* $(y, s)$ *respectively.*

For any $\varepsilon > 0$ and $z \in \mathbb{R}$, define

$$
\nu(\varepsilon, z) = \min\{A'(s) : |A(s) - A(z)| \le \varepsilon\}.
$$

Assume $A' \circ A^{-1}$ is Lipschitz, then we have the following elementary estimate which extends the result in [18, Corollary 6.4]:

$$
\begin{aligned}
|\partial_z [U_\varepsilon(z, k) - U_\varepsilon(k, z)]| &\le \frac{\varepsilon}{\nu(\varepsilon, z)} \mathcal{K}_1, \\
|\partial_z [F_\varepsilon(z, k) - F_\varepsilon(k, z)]| &\le \frac{\varepsilon}{\nu(\varepsilon, z)} \mathcal{K}_2,
\end{aligned}
\tag{3.10}
$$

where $k, z \in \mathbb{R}$, $\mathcal{K}_1 = \mathrm{L}(A' \circ A^{-1})$, $\mathcal{K}_2 = \mathcal{K}_1 \| f' \|_{L^\infty(\mathbb{R})} + \mathrm{L}(f')$ with $\mathrm{L}(A' \circ A^{-1})$ and $\mathrm{L}(f')$ being the Lipschitz constant of $A' \circ A^{-1}$ and $f'$ respectively.

To complete the Kružkov "doubling of variables" technique, we let first $l, m \to \infty$ then $n \to \infty$ in the entropy error inequality (3.9). The first two terms on the right-hand side of (3.9) can be treated by using (3.10) and the third and fourth terms can be shown to tend to zero. Thus we have

$$
\begin{aligned}
&- \int_Q U_\varepsilon(u, U_h) \theta_t + \int_Q H'_\varepsilon(A(u) - A(U_h)) |\nabla(A(U_h) - A(u))|^2 \theta \\
&\quad \le \mathcal{K} \varepsilon \int_Q \frac{1}{\nu(\varepsilon, U_h)} (|\partial_t U_h| + |\nabla_x U_h|) \theta - \int_0^T \langle \mathcal{R}, H_\varepsilon(A(U_h) - A(u)) \theta \rangle \, dt.
\end{aligned}
$$

where $\mathcal{K} = \max(\mathcal{K}_1, \mathcal{K}_2)$.

To proceed, we introduce the interior residual

$$
R^n := \bar{g}^n - \frac{U_h^n - \bar{U}_h^{n-1}}{\tau_n} + \Delta A(U_h^n) \quad \text{on any } K \in \mathcal{M}^n,
$$

where we recall that $\bar{g}^n = \tau_n^{-1} \int_{t^{n-1}}^{t^n} g(x, t) \, dt$.

**Theorem 3.6.** *Let $\varepsilon_0 = \sum_{i=1}^{3} \mathcal{E}_i$, where $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are the error indicators defined below. For any $m \geq 1$, let $Q_m = \Omega \times (0, t^m)$, and define*

$$\Lambda_m = \max\left(1, \int_{Q_m} \frac{1}{\nu(\varepsilon_0, U_h)}(|\partial_t U_h| + |\nabla U_h|) + \int_{\Omega} \frac{1}{\nu(\varepsilon_0, U_h^m)}\right), \quad (3.11)$$

*where for any $z \in \mathbb{R}$, $\nu(\varepsilon_0, z) = \min\{A'(s) : |A(s) - A(z)| \leq \varepsilon_0\}$. Then there exists a constant $C$ depending only on the minimum angles of the meshes $\mathcal{M}^n$, $n = 1, \ldots, m$, such that the following a posteriori error estimate is valid:*

$$\| u^m - U_h^m \|_{L^1(\Omega)} \leq \mathcal{E}_0 + \mathcal{E}_4 + \mathcal{E}_5 + C\Lambda_m^{1/2}\Big(\sum_{i=1}^{3} \mathcal{E}_i\Big),$$

*where the error indicators $\mathcal{E}_i$, $i = 0, \ldots, 5$, are defined by*

$$\mathcal{E}_0 = \| u_0 - U_h^0 \|_{L^1(\Omega)} \qquad \textit{initial error}$$

$$\mathcal{E}_1 = \Big(\sum_{n=1}^{m} \tau_n \| h_n^{1/2} [\![ \nabla A(U_h^n) ]\!] \|_{L^2(\Omega)}^2\Big)^{1/2} \qquad \textit{jump residual}$$

$$\mathcal{E}_2 = \Big(\sum_{n=1}^{m} \tau_n \| h_n R^n \|_{L^2(\Omega)}^2\Big)^{1/2} \qquad \textit{interior residual}$$

$$\mathcal{E}_3 = \Big(\sum_{n=1}^{m} \tau_n \| \nabla(A(U_h^n) - A(U_h^{n-1})) \|_{L^2(\Omega)}^2\Big)^{1/2} \qquad \textit{time residual}$$

$$\mathcal{E}_4 = \sum_{n=1}^{m} \int_{t^{n-1}}^{t^n} \left\| \frac{U_h^n - \bar{U}_h^{n-1}}{\tau_n} - (\partial_t U_h + \operatorname{div} f(U_h)) \right\|_{L^1(\Omega)} dt \qquad \textit{characteristic and coarsening}$$

$$\mathcal{E}_5 = \sum_{n=1}^{m} \int_{t^{n-1}}^{t^n} \| g - \bar{g}^n \|_{L^1(\Omega)}\, dt \qquad \textit{source.}$$

In the case of strong diffusion $A'(s) \geq \beta > 0$ for any $s \in \mathbb{R}$ and $A'$ is uniformly Lipschitz continuous, then $\Lambda_n$ is bounded by $\beta^{-1} \| U_h \|_{BV(Q_n)}$ which is expected to be bounded in practical computations. The a posteriori error estimator in Theorem 3.6 then recovers the standard a posteriori error estimator derived in the literature for parabolic problems [34], [9]. In particular, the space error indicators $\mathcal{E}_1^n$, $\mathcal{E}_2^n$, which control the adaptation of finite element meshes at each time step, are sharp in the sense that a local lower bound for the error can be established by extending the argument in [9, Theorem 2.2] for linear parabolic equations.

We also remark that the method of the a posteriori error analysis here is different from those for nonlinear conservation laws in [17], [18], [27] or nonlinear degenerate parabolic equations in [32]. Recall that there are several parameters introduced in the analysis:

- The regularizing parameter $\varepsilon$ in $H_\varepsilon(z)$.
- The boundary layer sequence parameters $\delta$ and $\eta$, and the mollifier parameters $l, m$ and $n$.

The analysis for Cauchy problems in [17], [18], [27] is based on letting $\varepsilon \to 0$ and taking finite mollifier parameters $l, m, n$. The analysis in [32] takes both finite $\varepsilon$ and finite mollifier parameters $l, m, n$. Note that there are no boundary layer sequence parameters $\delta, \eta$ for the analysis for Cauchy problems. The analysis in this paper is based on letting $\delta, \eta \to 0$ and $l, m, n \to \infty$ but taking a finite $\varepsilon$. We are not able to use the same technique as that in [17], [18], [27], [32] by choosing finite mollifier parameters $l, m, n$ to treat the problem with boundary conditions.

Based on the a posteriori error estimate in Theorem 3.6, an adaptive algorithm is proposed and implemented in [11]. In particular, the numerical experiments in [11] indicate that the total estimated error is roughly proportional to $M^{-1/3}$, i.e. $\eta \approx CM^{-1/3}$ for some constant $C > 0$. This implies the quasi-optimal decay of the error

$$\| u - U_h \|_{L^\infty(0,T;L^1(\Omega))} + \int_Q H'_\varepsilon(A(u) - A(U_h))|\nabla(A(U_h) - A(u))|^2 \le CM^{-1/3}$$

is valid asymptotically. Here $M$ is the sum of the number of elements of the space meshes over all time steps.

Figure 3.1 shows the meshes and the surface plots of the solutions at time $t = 0.251278$ and $t = 0.500878$ for the Burger's equation with small viscosity

$$\frac{\partial u}{\partial t} + u\partial_x u - \varepsilon \Delta u = 0 \quad \text{in } Q,$$

where $\Omega = (0, 1)^2$, $T = 1.0$, $\varepsilon = 10^{-3}$, and the initial condition and boundary condition

$$u(x, y, t)|_{\partial\Omega} = u_0(x, y) = 0.5 \sin(\pi x) + \sin(2\pi x).$$

The adaptive algorithm is based on the a posteriori error estimate in Theorem 3.6 and is described in [11]. We observe from Figure 3.1 that the method captures the internal and boundary layers of the solution.

**Acknowledgment.** The author would like to thank Shibin Dai, Guanghua Ji, Feng Jia, Xuezhe Liu, Ricardo H. Nochetto, Alfred Schmidt, and Haijun Wu for the joint work through the years.

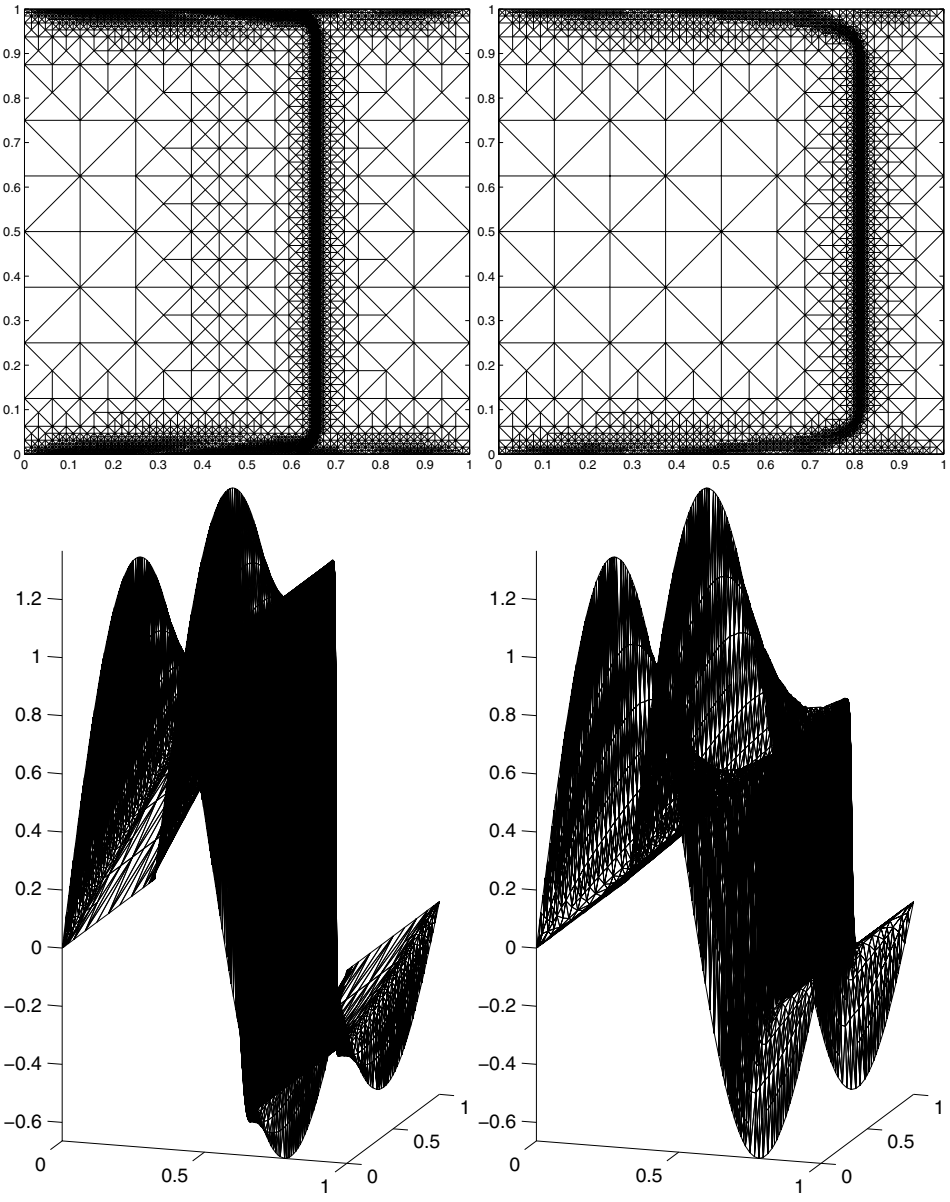


Figure 3.1. The meshes (top) and the surface plots (bottom) of the solutions $t = 0.400317$ (left) and $t = 1.0$ (right) with 35286 and 5020 nodes.

## References


[1] Alt, H. W., and Luckhaus, S., Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183** (1983), 311–341.

[2] Babuška, I., and Miller, A., A feedback finite element method with a posteriori error estimation: Part I. The finite element method and some basic properties of the a posteriori error estimator. *Comput. Meth. Appl. Mech. Engrg.* **61** (1987), 1–40.

[3] Babuška, I., and Rheinboldt, C., Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.* **15** (1978), 736–754.

[4] Berenger, J.-P., A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Physics* **114** (1994), 185–200.

[5] Binev, P., Dahmen, W., and DeVore, R., Adaptive finite element methods with convergence rates. *Numer. Math.* **97** (2004), 219–268.

[6] Carrillo, J., Entropy solutions for nonlinear degenerate problems. *Arch. Rational Mech. Anal.* **147** (1999), 269–361.

[7] Chen, Z., and Dai, S., Adaptive Galerkin methods with error control for a dynamical Ginzburg-Landau model in superconductivity. *SIAM J. Numer. Anal.* **38** (2001), 1961–1985.

[8] Chen, Z., and Dai, S., On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients. *SIAM J. Sci. Comput.* **24** (2002), 443–462.

[9] Chen, Z., and Jia, F., An adaptive finite element method with reliable and efficient error control for linear parabolic problems. *Math. Comp.* **73** (2004), 1163–1197.

[10] Chen, Z., and Ji, G., Adaptive computation for convection dominated diffusion problems. *Sci. China Ser. A* **47** (Supplement) (2004), 22–31.

[11] Chen, Z., and Ji, G., Sharp $L^1$ a posteriori error analysis for nonlinear convection-diffusion problems. *Math. Comp.* **75** (2006), 43–71.

[12] Chen, Z., and Liu, X., An Adaptive Perfectly Matched Layer Technique for Time-harmonic Scattering Problems. *SIAM J. Numer. Anal.* **43** (2005), 645–671.

[13] Chen, Z., and Nochetto, R. H., Residual type a posteriori error estimates for elliptic obstacke problems. *Numer. Math.* **84** (2000), 527–548.

[14] Chen, Z., Nochetto, R. H., and Schmidt, A., A characteristic Galerkin method with adaptive error control for continuous casting problem. *Comput. Methods Appl. Mech. Engrg.* **189** (2000), 249–276.

[15] Chen, Z., Nochetto, R. H., and Schmidt, A., Error control and adaptivity for a phase relaxation model. *Math. Model. Numer. Anal.* **34** (2000), 775–797.

[16] Chen, Z., and Wu, H., An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures. *SIAM J. Numer. Anal.* **41** (2003), 799–826.

[17] Cockburn, B., Coquel, B. F., and Lefloch, P. G., An error estimate for finite volume methods for multidimensional conservation laws. *Math. Comp.* **63** (1994), 77–103.

[18] Cockburn, B., and Gremaud, P.-A., Error estimates for finite element methods for scalar conservation laws. *SIAM J. Numer. Anal.* **33** (1996), 522–554.

[19] Collino, F., and Monk, P. B., The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comput.* **19** (1998), 2061–2090.

[20] Colton D., and Kress R., *Integral Equation Methods in Scattering Theory.* John Wiley & Sons, New York 1983.

[21] Hohage, T., Schmidt, F., and Zschiedrich, L., Solving time-harmonic scattering problems based on the pole condition. II: Convergence of the PML method. *SIAM J. Math. Anal.*, to appear.

[22] Lassas, M., and Somersalo, E., On the existence and convergence of the solution of PML equations. *Computing* **60** (1998), 229–241.

[23] Dörfler, W., A convergent adaptive algorithm for Possion's equations. *SIAM J. Numer. Anal.* **33** (1996), 1106–1124.

[24] Douglas Jr., J., and Russell, T. F., Numerical methods for convection-dominated diffusion problem based on combining the method of characteristic with finite element or finite difference procedures. *SIAM J. Numer. Anal.* **19** (1982), 871–885.

[25] Eriksson, K., and Johnson, C., Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM J. Numer. Anal.* **28** (1991), 43–77.

[26] Houston, P., and Süli, E., Adaptive Lagrange-Galerkin methods for unsteady convection-diffusion problems. *Math. Comp.* **70** (2000), 77–106.

[27] Kröner, D., and Ohlberger, M., A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi-dimensions. *Math. Comp.* **69** (2000), 25–39.

[28] Kružkov, N. N., First order quasi-linear equations in several independent variables. *Math. USSR Sb.* **10** (1970), 217–243.

[29] Mascia, C., Porretta, A. and Terracina, A., Nonhomogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations. *Arch. Rational Mech. Anal.* **163** (2002), 87–124.

[30] Morin, P., Nochetto. R. H., and Siebert, K. G., Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.* **38** (2000), 466–488.

[31] Nochetto, R. H., Schmidt, A., and Verdi, C., A posteriori error estimation and adaptivity for degenerate parabolic problems. *Math. Comp.* **69** (2000), 1–24.

[32] Ohlberger, M., A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations. *Math. Model. Numer. Anal.* **35** (2001), 355–387.

[33] Otto, F., $L^1$-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Diff. Equations* **131** (1996), 20–38.

[34] Picasso, M., Adaptive finite elements for a linear parabolic problem. *Comput. Methods Appl. Mech. Engrg.* **167** (1998), 223–237.

[35] Pironneau, O., On the transport-diffusion algorithm and its application to the Navier-Stokes equations. *Numer. Math.* **38** (1982), 309–332.

[36] Schmidt, A., and Siebert, K. G., ALBERT: An adaptive hierarchical finite element toolbox. *IAM, University of Freiburg,* 2000; http://www.mathematik.uni-freiburg.de/IAM/Research/projectsdz/albert.

[37] Teixeira, F. L., and Chew, W. C., Advances in the theory of perfectly matched layers. In *Fast and Efficient Algorithms in Computational Electromagnetics* (ed. by W. C. Chew), Artech House, Boston 2001, 283–346.

[38] Turkel, E., and Yefet, A., Absorbing PML boundary layers for wave-like equations. *Appl. Numer. Math.* **27** (1998), 533–557.

[39] Verfürth, R., *A Review of A Posteriori Error Estimation and adaptive Mesh Refinement Techniques*. Advances in Numerical Mathematics, John Wiley & Sons, B. G. Teubner, Chichester, Stuttgart 1996.

LSEC, Institute of Computational Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

E-mail: zmchen@lsec.cc.ac.cn

# Error estimates for anisotropic finite elements and applications

Ricardo G. Durán*

**Abstract.** The finite element method is one of the most frequently used techniques to approximate the solution of partial differential equations. It consists in approximating the unknown solution by functions which are polynomials on each element of a given partition of the domain, made of triangles or quadrilaterals (or their generalizations to higher dimensions).

A fundamental problem is to estimate the error between the exact solution $u$ and its computable finite element approximation. In many situations this error can be bounded in terms of the best approximation of $u$ by functions in the finite element space of piecewise polynomial functions. A natural way to estimate this best approximation is by means of the Lagrange interpolation or other similar procedures.

Many works have considered the problem of interpolation error estimates. The classical error analysis for interpolations is based on the so-called *regularity assumption*, which excludes elements with different sizes in each direction (called *anisotropic*). The goal of this paper is to present a different approach which has been developed by many authors and can be applied to obtain error estimates for several interpolations under more general hypotheses.

An important case in which anisotropic elements arise naturally is in the approximation of convection-diffusion problems which present boundary layers. We present some applications to these problems.

Finally we consider the finite element approximation of the Stokes equations and present some results for non-conforming methods.

**Mathematics Subject Classification (2000).** Primary 65N30; Secondary 65N15.

**Keywords.** Finite elements, Mixed methods, anisotropic elements, Stokes equations, convection-diffusion.

## 1. Introduction

The finite element method in its different variants is one of the most frequently used techniques to approximate the solution of partial differential equations. The general idea is to use weak or variational formulations in an infinite dimensional space and to replace that space by a finite dimensional one made of piecewise polynomial functions. In this way, the original differential equation is transformed into an algebraic problem which can be solved by computational methods. Although the main idea goes back

*Supported by ANPCyT under grant PICT 03-05009, by Universidad de Buenos Aires under grant X052 and by Fundación Antorchas. The author is a member of CONICET, Argentina.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

to the works of Galerkin and Ritz in the early twentieth-century (or even to previous works, see for example [9] for a discussion of the history of these ideas), the finite element method became more popular since the middle of the twentieth century mainly because of its application by engineers to structural mechanics. On the other hand, the general mathematical analysis started only around forty years ago.

The theory of finite elements can be divided into *a priori* and *a posteriori* error analysis. The main goals of the *a priori* analysis are to prove convergence of the methods, to know the order of convergence (in terms of parameters associated with the finite dimensional problem, such as degree of approximation, mesh-size, size of the discrete problem, geometry of the elements, etc.) and the dependence of the error on properties of the unknown exact solution (such as its smoothness, which in many cases is already known from the theory of partial differential equations). Instead, the goals of the *a posteriori* error analysis are to obtain more quantitative information on the error and to develop self-adaptive methods to improve the approximation iteratively.

In this paper we consider several problems related to *a priori* error estimates. We will deal mainly with the error analysis for flat or anisotropic elements, which arise naturally in several applications.

Let us begin by recalling the basic ideas of weak formulations of differential equations and finite element approximations. A general abstract formulation for linear problems is given by

$$B(u, v) = F(v) \quad \text{for all } v \in V, \tag{1.1}$$

where $u \in V$ is the solution to be found, $V$ is a Hilbert space, $F$ is a continuous linear form and $B$ is a continuous bilinear form, i.e., there exists a constant $M > 0$ such that

$$|B(u, v)| \leq M\|u\|\|v\|$$

where $\|\cdot\|$ is the norm in the Hilbert space $V$.

To approximate the solution, we want to introduce a finite dimensional space $V_h \subset V$. The usual way to do this is to introduce a partition $\mathcal{T}_h$ of the domain $\Omega$ where we want to solve the differential equation usually made of triangular or quadrilateral elements (or their generalizations in 3D). The parameter $h$ is usually related to the mesh size. Then the space $V_h$ consists of functions which restricted to each element of the partition are polynomials.

The approximate solution of our problem is $u_h \in V_h$ that satisfies

$$B(u_h, v) = F(v) \quad \text{for all } v \in V_h.$$

Assume that the form $B$ is coercive, namely, that there exists a constant $\alpha > 0$ such that

$$B(v, v) \geq \alpha\|v\|^2 \quad \text{for all } v \in V. \tag{1.2}$$

Then the classical error analysis is based on Cea's lemma (see [14]), which states that

$$\|u - u_h\| \leq \frac{M}{\alpha}\|u - v\| \quad \text{for all } v \in V_h. \tag{1.3}$$

Notice that (1.2) also guarantees existence and uniqueness of solution in $V$ as well as in $V_h$, thanks to the well-known Lax–Milgram theorem.

If this condition does not hold, but the form $B$ satisfies the so-called *inf-sup conditions*, that is, there exists $\beta > 0$ such that

$$\inf_{u\in V_h} \sup_{v\in V_h} \frac{B(u,v)}{\|u\|\|v\|} \geq \beta, \tag{1.4}$$

$$\inf_{v\in V_h} \sup_{u\in V_h} \frac{B(u,v)}{\|u\|\|v\|} \geq \beta, \tag{1.5}$$

then we also have

$$\|u-u_h\| \leq \frac{M}{\beta}\|u-v\| \quad \text{for all } v\in V_h. \tag{1.6}$$

If the above inf-sup conditions hold in $V$, we also have uniqueness and existence of solution. However, this is not sufficient to obtain (1.6), as the inf-sup conditions are not inherited by subspaces. This is the main difference between error analysis of coercive and non-coercive forms which satisfy (1.4) and (1.5).

The classical example of a form $B$ which satisfies the inf-sup conditions but is not coercive, is the form associated to the Stokes equations of fluid dynamics (see for example [13], [20]).

In view of (1.3) and (1.6), in order to obtain an estimate for $\|u-u_h\|$ it is enough to bound $\|u-v\|$ for a function $v\in V_h$. Therefore this is one of the most important problems in the theory of finite elements. Usually, the function $v$ is taken to be a Lagrange interpolation of $u$. However, in some cases it is more convenient to use different approximations.

In many problems it is convenient to use spaces $V_h$ which are not contained in $V$. These methods are called non-conforming and in this case the right-hand sides of (1.3) and (1.6) are modified by adding the so-called "consistency terms". One of the best-known methods of this type is that of Crouzeix–Raviart, which is closely related to the mixed finite element methods of Raviart–Thomas (see [8], [23]).

The goal of this paper is to present general ideas to obtain error estimates for different interpolations valid under very general hypotheses on the elements, in particular, allowing meshes with flat or anisotropic elements. We consider Lagrange and other kind of interpolations arising in mixed finite element methods and give some applications to the approximation of convection-diffusion equations for which anisotropic elements are needed due to the presence of boundary layers.

Finally we consider the finite element approximation of the Stokes equations and recall some results for non-conforming methods.

## 2. Notation and some basic inequalities

The classical finite element analysis for triangular elements requires the so-called regularity assumption, i.e.,

$$\frac{h_T}{\rho_T} \leq C \tag{2.1}$$

where $h_T$ and $\rho_T$ are the outer and inner diameter, respectively (see Figure 1). In other words, the constants in the error estimates depend on $C$ (see for example [11], [14]).

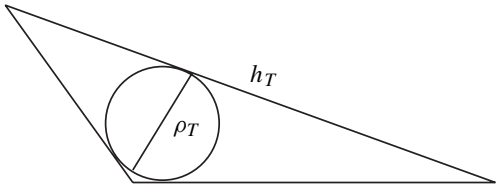



Figure 1

The same hypothesis is also needed for the analysis of mixed and non-conforming methods (see [15] and [24]).

For standard Lagrange interpolation on conforming elements, since the works of Babuska–Azis [10] and Jamet [21] it is well known that the regularity assumption can be relaxed. For example, in the case of triangles it can be replaced by the weaker maximum angle condition (i.e. angles bounded away from $\pi$). For rectangular elements, optimal error estimates can be obtained for arbitrary rectangles (while the regularity assumption requires that the edge sizes be comparable). In the case of general quadrilaterals, the situation is more complicated and several conditions, weaker than regularity, have been introduced to prove the error estimates (see, for example, [3]).

The standard method to prove error estimates is to obtain them first in a reference element and then to make a change of variables (see [14]). A different approach is to work directly in a given element and to use Poincaré type inequalities. The main idea is that the interpolation error usually has some vanishing averages (on the element, or edges, or faces, depending of the kind of interpolation considered). In this approach, the reference element is sometimes used to obtain the Poincaré type inequalities but, since one is bounding an $L^2$-norm, the constants appearing in the estimates are independent of the aspect ratio of the element.

We will use the following notation. By $H^1(\Omega)$ we mean the usual Sobolev space of $L^2$ functions with distributional first derivatives in $L^2$ and by $H_0^1(\Omega)$ the subspace of $H^1(\Omega)$ of functions vanishing on the boundary.

Similarly, $W^{k,p}(\Omega)$, for $1 \leq p \leq \infty$, indicates the Sobolev space of $L^p(\Omega)$ functions with distributional derivatives of order $k$ in $L^p(\Omega)$. When $p = 2$ we set $H^k(\Omega) = W^{k,2}(\Omega)$.

Here $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, is a bounded domain. For a general triangle $T$, $h_T$ is its diameter, $\boldsymbol{p}_0$ is a vertex (arbitrary unless otherwise specified), $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ (with $\|\boldsymbol{v}_i\| = 1$) are the directions of the edges $\ell_1$, $\ell_2$ sharing $\boldsymbol{p}_0$ (see Figure 2), and $\nu_i$ is the exterior unit normal to the side $\ell_i$ (with obvious generalizations to 3D). We also use the standard notation $\mathcal{P}_k$ for polynomials of total degree less than or equal to $k$, and $\mathcal{Q}_k$ for polynomials of degree less than or equal to $k$ in each variable. We call $\hat{T}$ the reference triangle with vertices at $(0, 0)$, $(1, 0)$ and $(0, 1)$, and $F\colon \hat{T} \to T$ the affine transformation $F(\hat{x}) = B\hat{x} + \boldsymbol{p}_0$ with $B\boldsymbol{e}_i = l_i \boldsymbol{v}_i$, where $\boldsymbol{e}_i$ are the canonical vectors.

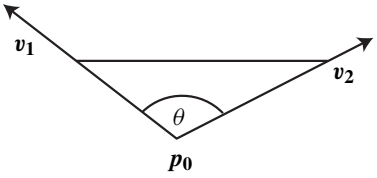



Figure 2

The following two results are the classical Poincaré inequality and a generalization of it (first given in [10]) written in a convenient way for our purposes.

**Lemma 2.1.** *Let $T$ be a triangle (resp. tetrahedron) and let $f \in H^1(T)$ be a function with vanishing average on $T$. Then there exists a constant $C$ independent of $T$ and of $f$ such that*

$$\|f\|_{L^2(T)} \leq C \sum_{j=1}^{n} |\ell_j| \left\| \frac{\partial f}{\partial \boldsymbol{v}_j} \right\|_{L^2(T)}. \tag{2.2}$$

*Proof.* It follows from the Poincaré inequality on $\hat{T}$ and making the change of variables $F$. □

**Lemma 2.2.** *Let $T$ be a triangle (resp. tetrahedron) and $\ell$ be any of its edges (resp. faces). Let $f \in H^1(T)$ be a function with vanishing average on $\ell$. Then there exists a constant $C$ independent of $T$ such that*

$$\|f\|_{L^2(T)} \leq C \sum_{j=1}^{n} |\ell_j| \left\| \frac{\partial f}{\partial \boldsymbol{v}_j} \right\|_{L^2(T)}. \tag{2.3}$$

*Proof.* It is enough to prove that, on the reference element $\hat{T}$,

$$\|f\|_{L^2(\hat{T})} \leq C \|\nabla f\|_{L^2(\hat{T})}. \tag{2.4}$$

Then, for a general triangle, the result follows by making the change of variables $F$.

The estimate (2.4) can be proved by a standard compactness argument (as was done in [10]). A different proof can be given by a using (2.2) and a trace theorem.

Indeed, if $f_{\hat{\ell}}$ and $f_{\hat{T}}$ denote the averages on $\hat{\ell}$ and $\hat{T}$, respectively, and if we assume that $f_{\hat{\ell}} = 0$ we have

$$\|f\|_{L^2(\hat{T})} = \|f - f_{\hat{\ell}}\|_{L^2(\hat{T})} \leq \|f - f_{\hat{T}}\|_{L^2(\hat{T})} + \|f_{\hat{T}} - f_{\hat{\ell}}\|_{L^2(\hat{T})}.$$

But

$$f_{\hat{T}} - f_{\hat{\ell}} = \frac{1}{|\hat{\ell}|}\int_{\hat{\ell}}(f_{\hat{T}} - f),$$

and therefore an application of a standard trace theorem gives

$$\|f_{\hat{T}} - f_{\hat{\ell}}\|_{L^2(\hat{T})} \leq C\{\|f - f_{\hat{T}}\|_{L^2(\hat{T})} + \|\nabla f\|_{L^2(\hat{T})}\}$$

with a constant $C$ which depends only on the reference element. Hence (2.4) follows from (2.2). □

## 3. Error estimates for Lagrange interpolation

**3.1. The two-dimensional case.** To introduce the general idea we present first two simple classical cases: the Lagrange interpolation for lowest degree finite elements in triangles or rectangles. The argument is essentially that given in [10] for triangles. In the case of rectangles, an extra step is required due to the presence of a non-vanishing second derivative of the interpolating function.

Given a triangle $T$ we denote with $I_1 u \in \mathcal{P}_1$ the Lagrange interpolation of $u$, i.e., the affine function which equals $u$ on the vertices of $T$. $D^2 u$ denotes the sum of the absolute values of second derivatives of $u$.

**Theorem 3.1.** *There exists a constant $C$ such that, if $\theta$ is the maximum angle of $T$,*

$$\|\nabla(u - I_1 u)\|_{L^2(T)} \leq \frac{C}{\sin\theta}\, h_T \|D^2 u\|_{L^2(T)}.$$

*Proof.* Observe that, for $i = 1, 2$, $\nabla(u - I_1 u)\cdot \boldsymbol{v}_i$, has vanishing average on one side of $T$. Therefore, applying Lemma 2.2 and using that the second derivatives of $I_1 u$ vanish, we obtain

$$\|\nabla(u - I_1 u)\cdot \boldsymbol{v}_i\|_{L^2(T)} \leq C\left\{|\ell_1|\left\|\frac{\partial \nabla u\cdot \boldsymbol{v}_i}{\partial \boldsymbol{v}_1}\right\|_{L^2(T)} + |\ell_2|\left\|\frac{\partial \nabla u\cdot \boldsymbol{v}_i}{\partial \boldsymbol{v}_2}\right\|_{L^2(T)}\right\}.$$

Then, if we choose $\boldsymbol{p}_0$ as the vertex corresponding to the maximum angle of $T$, we have

$$|\nabla(u - I_1 u)| \leq \frac{C}{\sin\theta}\{|\nabla(u - I_1 u)\cdot \boldsymbol{v}_1| + |\nabla(u - I_1 u)\cdot \boldsymbol{v}_2|\},$$

and hence the theorem is proved. □

We consider now the case of rectangles. We use the same notation, $I_1 u$, for the interpolation which now belongs to $\mathcal{Q}_1$. The proof for this case is analogous to the previous one, with the only difference that $\frac{\partial^2 I_1 u}{\partial x \partial y}$ does not vanish.

Let $R$ be a rectangle and let $\ell_1$, $\ell_2$ be two adjacent sides. Clearly, the result of Lemma 2.2 holds for this case also.

**Theorem 3.2.** *There exists a constant $C$, independent of the relation between $|\ell_1|$ and $|\ell_1|$, such that*

$$\left\| \frac{\partial}{\partial x}(u - I_1 u) \right\|_{L^2(R)} \leq C \left\{ |\ell_1| \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{L^2(R)} + |\ell_2| \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{L^2(R)} \right\} \tag{3.1}$$

*and*

$$\left\| \frac{\partial}{\partial y}(u - I_1 u) \right\|_{L^2(R)} \leq C \left\{ |\ell_1| \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{L^2(R)} + |\ell_2| \left\| \frac{\partial^2 u}{\partial y^2} \right\|_{L^2(R)} \right\}. \tag{3.2}$$

*Proof.* Proceeding as in the case of triangles, we have

$$\left\| \frac{\partial}{\partial x}(u - I_1 u) \right\|_{L^2(R)} \leq C \left\{ |\ell_1| \left\| \frac{\partial^2 (u - I_1 u)}{\partial x^2} \right\|_{L^2(R)} + |\ell_2| \left\| \frac{\partial^2 (u - I_1 u)}{\partial x \partial y} \right\|_{L^2(R)} \right\}. \tag{3.3}$$

But, $\frac{\partial^2 I_1 u}{\partial x^2} = 0$ and an elementary computation shows that

$$\int_R \frac{\partial^2 I_1 u}{\partial x \partial y} = \int_R \frac{\partial^2 u}{\partial x \partial y},$$

i.e., $\frac{\partial^2 I_1 u}{\partial x \partial y}$ is the average of $\frac{\partial^2 u}{\partial x \partial y}$ on $R$. Then

$$\left\| \frac{\partial^2 I_1 u}{\partial x \partial y} \right\|_{L^2(R)} \leq \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{L^2(R)}$$

and therefore (3.1) holds. Obviously, the proof of (3.2) is analogous. □

**Remark 3.3.** If the function $u \in H^3(R)$, then the last term on the right-hand side of (3.3) is of higher order. Indeed, that term is the difference between $\frac{\partial^2 u}{\partial x \partial y}$ and its average. Therefore we have the estimate

$$\left\| \frac{\partial}{\partial x}(u - I_1 u) \right\|_{L^2(R)} \leq C |\ell_1| \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{L^2(R)} + \text{ higher order terms.}$$

**3.2. The three-dimensional case.** Many results on finite elements can be extended almost straightforward from 2D to 3D. However, this is not the case of error estimates for anisotropic elements. Indeed, counterexamples for an estimate analogous to (3.1) in the 3D case have been given in [6] and [26]. They show that the constant in the

estimate blows-up when a rectangular reference tetrahedron (or cube) is compressed in one direction.

Many papers have been published considering the 3D case. For example, in the case of tetrahedra, Krízek [22] introduced a natural generalization of the maximum angle condition: if the angles between faces and the angles in the faces are bounded away from $\pi$, he obtained error estimates for smooth functions, namely, $u \in W^{2,\infty}$. In [16] the results of Krízek were extended to functions in $W^{2,p}$ with $2 < p < \infty$ (and, moreover, to functions in an intermediate Orlicz space between $H^2$ and $W^{2,p}$, $p > 2$). Therefore, although the estimate fails for functions in $H^2$, it is valid for functions only slightly more regular. Let us mention that the reason why the arguments applied in 2D cannot be generalized, is that the estimate given in Lemma 2.2 is not true in 3D if $\ell$ is an edge instead of a face (note that the interpolation error for the Lagrange interpolation has vanishing integral on edges).

On the other hand, many papers have considered error estimates for different interpolations (see for example [1], [5], [16], [17]), namely, different variants of average interpolators. This kind of interpolations have been introduced to approximate non-smooth functions (for which the Lagrange interpolation is not even defined). However, they have as well better approximation properties on anisotropic elements for functions in $H^2$. Indeed, using average interpolations, the 2D results can be generalized to 3D. Observe that, in view of (1.3) and (1.6), error estimates for an average interpolation will give bounds for finite element approximations.

## 4. Applications to convection-diffusion equations

A very important application in which anisotropic elements are needed is the approximation of convection-diffusion problems in which boundary layers arise.

Consider for example the model problem

$$\begin{aligned} -\varepsilon \Delta u + b \cdot \nabla u + cu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $\Omega = (0, 1)^2$ and $\varepsilon > 0$ is a small parameter.

It is well known that the numerical approximation of this equation requires some special method in order to obtain good results when the problem is convection dominated, due to the presence of boundary or interior layers. In the case of boundary layers, one possibility is to use appropriate refined meshes near the boundary; this methodology gives rise to anisotropic elements. Using estimates (3.1) and (3.2) it is possible to obtain quasi-optimal order convergence (with respect to the number of nodes) in the $\varepsilon$-norm defined by

$$\|v\|_\varepsilon^2 = \|v\|_{L^2(\Omega)}^2 + \varepsilon \|\nabla v\|_{L^2(\Omega)}^2$$

for the standard $\mathcal{Q}_1$ approximation on appropriate graded meshes.

This problem can be written in the general form (1.1) with $V = H_0^1(\Omega)$,

$$B(u, v) = \int_\Omega (\varepsilon \nabla u \nabla v + b \cdot \nabla u\, v + c\, uv)\ dx$$

and

$$F(v) = \int_\Omega f v\, dx.$$

Assuming that there exists a constant $\mu$ independent of $\varepsilon$ such that

$$c - \frac{\operatorname{div} b}{2} \geq \mu > 0, \tag{4.2}$$

the bilinear form $B$ is coercive in the $\varepsilon$-norm uniformly in $\varepsilon$ (see [25]), i.e., the constant $\alpha$ in (1.2) is independent of $\varepsilon$. However, the continuity of $B$ is not uniform in $\varepsilon$ and this is one of the reasons why it is not possible to apply directly the general result (1.3) to obtain error estimates valid uniformly in $\varepsilon$. Therefore, a special analysis is required and this was the object of [18]. It was proved in that paper that

$$\|u - u_h\|_\varepsilon \leq C \frac{\log^2(1/\varepsilon)}{\sqrt{N}},$$

where $N$ is the number of nodes and $h > 0$ is a parameter associated with the meshes. Observe that this order of convergence is quasi-optimal in the sense that, up to the logarithm factor, it is the same order that one obtains for a smooth solution of a problem with $\varepsilon = O(1)$ using uniform meshes.

Assuming that the coefficient $b$ is such that the boundary layers are close to $x = 0$ and $y = 0$, the meshes $\mathcal{T}_h$ are such that the grading in each direction is given by

$$\begin{cases} \xi_0 = 0, & \\ \xi_i = ih\varepsilon & \text{for } 1 \leq i < \frac{1}{h} + 1, \\ \xi_{i+1} = \xi_i + h\xi_i & \text{for } \frac{1}{h} + 1 \leq i \leq M - 2, \\ \xi_M = 1, & \end{cases} \tag{4.3}$$

where $M$ is such that $\xi_{M-1} < 1$ and $\xi_{M-1} + h\xi_{M-1} \geq 1$. We assume that the last interval $(\xi_{M-1}, 1)$ is not too small in comparison with the previous one $(\xi_{M-2}, \xi_{M-1})$ (if this is not the case, we just eliminate the node $\xi_{M-1}$).

Figure 3 shows the approximate solution of (4.1) for

$$\varepsilon = 10^{-6}, \quad b = (1 - 2\varepsilon)(-1, -1), \quad c = 2(1 - \varepsilon)$$

and

$$f(x, y) = -\left[x - \left(\frac{1 - e^{-\frac{x}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}}\right) + y - \left(\frac{1 - e^{-\frac{y}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}}\right)\right] e^{x+y}.$$

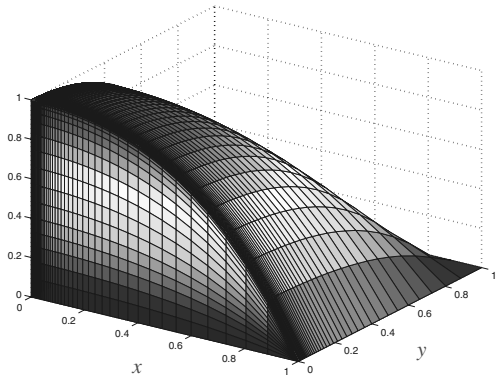



Figure 3

Observe that no oscillations arise although we are using the standard $\mathcal{Q}_1$ finite element method.

The graded meshes are an alternative to the well-known Shishkin meshes which have been widely analyzed for convection-diffusion problems (see for example [25]).

From the error analysis given in [18] one can see that a graded mesh designed for a value of $\varepsilon$ works well also for larger values of $\varepsilon$. This is not the case for Shishkin meshes. Table 1 shows the values of the $\varepsilon$-norm of the error for different values of $\varepsilon$, solving the problem with the mesh corresponding to $\varepsilon = 10^{-6}$, using graded meshes and Shishkin meshes.

Table 1

| $\varepsilon$ | Error |
|---|---|
| $10^{-6}$ | 0.040687 |
| $10^{-5}$ | 0.033103 |
| $10^{-4}$ | 0.028635 |
| $10^{-3}$ | 0.024859 |
| $10^{-2}$ | 0.02247 |
| $10^{-1}$ | 0.027278 |

Graded meshes, $N = 10404$.

| $\varepsilon$ | Error |
|---|---|
| $10^{-6}$ | 0.0404236 |
| $10^{-5}$ | 0.249139 |
| $10^{-4}$ | 0.623650 |
| $10^{-3}$ | 0.718135 |
| $10^{-2}$ | 0.384051 |
| $10^{-1}$ | 0.0331733 |

Shishkin meshes, $N = 10609$.

To see the different structures, we show in Figure 4 a Shishkin mesh (on the right) and one of our graded meshes (on the left) having the same number of nodes. For the sake of clarity, we show only the part of the meshes corresponding to $(0, 1/2) \times (0, 1/2)$ and $\varepsilon = 10^{-\frac{3}{2}}$.

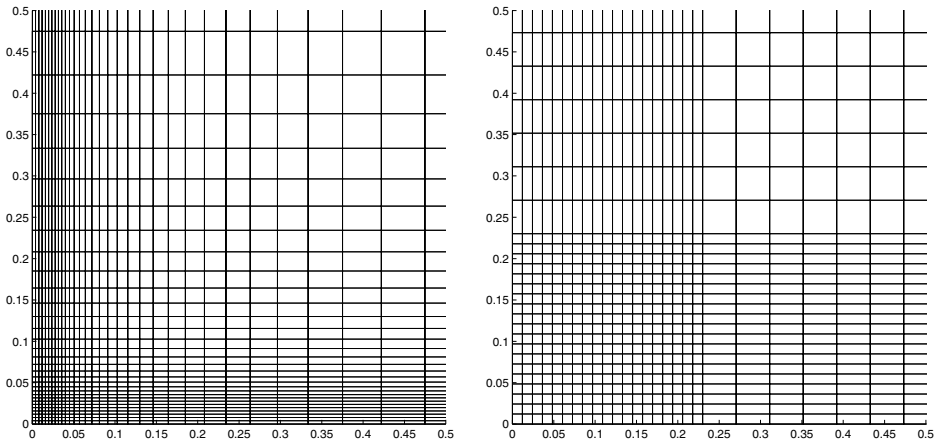



Figure 4

## 5. Error estimates for Raviart–Thomas interpolation

**5.1. The two dimensional case.** The Raviart–Thomas spaces were introduced in [24] to approximate vector fields $\boldsymbol{u} \in H(\mathrm{div}, \Omega)$ where

$$H(\mathrm{div}, \Omega) = \{\boldsymbol{u} \in L^2(\Omega) : \mathrm{div}\, \boldsymbol{u} \in L^2\}.$$

For any integer $k \geq 0$, the space $\mathcal{RT}_k$ on a triangle $T$ is defined by

$$\mathcal{RT}_k(T) = \mathcal{P}_k^2(T) \oplus (x, y)\mathcal{P}_k(T).$$

Calling $P_k$ the $L^2$ orthogonal projection on $\mathcal{P}_k(T)$, it is known (see [24]) that there exists an operator $RT_k \colon H^1(T)^2 \to \mathcal{RT}_k(T)$ satisfying the following commutative diagram property:

$$\begin{array}{ccccc} H^1(T)^2 & \xrightarrow{\mathrm{div}} & L^2(T) & & \\ \Big\downarrow{\scriptstyle RT_k} & & \Big\downarrow{\scriptstyle P_k} & & \\ \mathcal{RT}_k(T) & \xrightarrow[\mathrm{div}]{} & \mathcal{P}_k(T) & \longrightarrow & 0. \end{array} \tag{5.1}$$

For the case of anisotropic elements, only the lowest degree case $\mathcal{RT}_0$ has been considered. Error estimates for this case have been obtained in [2].

Below we will show how the arguments can be generalized to obtain error estimates for the case of $\mathcal{RT}_1$. Higher order approximations can be treated similarly although this extension is not straightforward.

Let us first recall the results for $\mathcal{RT}_0$. Again, the results follow by the generalized Poincaré inequality given in Lemma 2.2 as we show in the next theorem.

**Theorem 5.1.** *There exists a constant $C$ such that, if $\theta$ is the maximum angle of $T$,*

$$\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)} \leq \frac{C}{\sin\theta} \sum_{k=1}^{2} |\ell_k| \Big( \Big\| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}_k} \Big\|_{L^2(T)} + \|\operatorname{div} \boldsymbol{u}\|_{L^2(T)} \Big).$$

*Proof.* Since $(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i$ has zero mean value on $\ell_i$, it follows from Lemma 2.2 that

$$\|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \leq C \sum_{k=1}^{2} |\ell_k| \Big\| \frac{\partial (\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i}{\partial \boldsymbol{v}_k} \Big\|_{L^2(T)}. \tag{5.2}$$

But it is easy to check that

$$\frac{\partial (RT_0\boldsymbol{u} \cdot \nu_i)}{\partial \boldsymbol{v}_k} = \frac{1}{2} (\operatorname{div} RT_0\boldsymbol{u})\, \boldsymbol{v}_k \cdot \nu_i.$$

On the other hand, using the commutative diagram property (5.1), we have

$$\|\operatorname{div} RT_0\boldsymbol{u}\|_{L^2(T)} \leq \|\operatorname{div} \boldsymbol{u}\|_{L^2(T)}$$

and so it follows from (5.2) that

$$\|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \leq C \sum_{k=1}^{2} |\ell_k| \Big( \Big\| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}_k} \Big\|_{L^2(T)} + \|\operatorname{div} \boldsymbol{u}\|_{L^2(T)} |\nu_i \cdot \boldsymbol{v}_k| \Big). \tag{5.3}$$

Up to now the constant $C$ is independent of $T$. If we want to bound $\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)}$, it is natural to expect that the constant will depend on the geometry of the element.

In view of (5.3) it would be enough to control $\boldsymbol{u} - RT\boldsymbol{u}$ in terms of its components in the directions of the normals to the edges. For a fixed triangle the estimate

$$|\boldsymbol{u} - RT_0\boldsymbol{u}| \leq C\{|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_1| + |(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_2|\}$$

holds. Moreover, for a family of triangles, the constant $C$ will not degenerate if the angle between $\nu_1$ and $\nu_2$ does not go to 0 or $\pi$ or, equivalently, if the angle between the corresponding edges does not go to 0 or $\pi$. Therefore the constant will be uniformly bounded for a family of elements with maximum angle bounded away from $\pi$. More precisely, we have

$$\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)} \leq \frac{C}{\sin\theta} \sum_{i=1}^{2} \|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \tag{5.4}$$

where $\theta$ is the maximum angle of $T$. Indeed, if $N$ is the matrix which has $\nu_1$ and $\nu_2$ as its rows, then

$$\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)} \leq \|N^{-1}\| \sum_{i=1}^{2} \|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)}$$

where $\|\cdot\|$ denotes the matrix norm associated with the euclidean norm. But since the $\nu_i$ are unit vectors it follows that $\|N^{-1}\| \leq \frac{C}{|\det N|}$ and $|\det N| = \sin\theta_1$, where $\theta_1$ is the angle between $\nu_1$ and $\nu_2$. If the vertex $\boldsymbol{p}_0$ is the one corresponding to the maximum angle $\theta_1 = \pi - \theta$, then (5.4) holds and the theorem is proved. □

Similar arguments can be applied for the analysis of higher order elements. However the extension is not straightforward. In what follows we consider the case of $\mathcal{RT}_1$. This case requires the following generalization of the Poincaré inequality.

**Lemma 5.2.** *Let $T$ be a triangle and $\ell$ one of its sides. If $f \in H^2(T)$ satisfies*

$$\int_\ell fp = 0 \quad \textit{for all } p \in \mathcal{P}_1(\ell) \quad \textit{and} \quad \int_T f = 0,$$

*then*

$$\|f\|_{L^2(T)} \leq Ch_T^2 \|D^2 f\|_{L^2(T)}$$

*with a constant $C$ independent of the shape of the triangle.*

*Proof.* Observing that, if $f \in \mathcal{P}_1$ satisfies the three hypotheses of the lemma then $f = 0$, it follows by standard compactness arguments that

$$\|f\|_{L^2(\hat{T})} \leq C\|D^2 f\|_{L^2(\hat{T})}.$$

Then an affine change of variables concludes the proof. □

To obtain the error estimate for the $RT_1$ interpolation we will need to have a bound for the gradient of the $P_1$ projection. This is the goal of the next lemma.

**Lemma 5.3.** *If $f \in H^1(T)$ we have*

$$\|\nabla P_1 f\|_{L^2(T)} \leq C\|\nabla f\|_{L^2(T)}$$

*with a constant $C$ depending only on the maximum angle of $T$.*

*Proof.* We will prove that for the triangle with vertices at $(0, 0)$, $(h, 0)$ and $(0, 1)$ we have

$$\|\nabla P_1 f\|_{L^2(T)} \leq 6\|\nabla f\|_{L^2(T)}.$$

Then the general result follows by an affine change of variables.

Let $M_i, i = 1, 2, 3$ be the mid-side points of $T$. Since the quadrature rule obtained by interpolating at these points is exact for quadratic polynomials, it is easy to see that the functions

$$\phi_1 = \left(\frac{6}{h}\right)^{1/2}(1-2y), \quad \phi_2 = \left(\frac{6}{h}\right)^{1/2}\left(2y+\frac{2x}{h}-1\right) \quad \text{and} \quad \phi_3 = \left(\frac{6}{h}\right)^{1/2}\left(1-\frac{2x}{h}\right)$$

form an orthonormal basis of $\mathcal{P}_1(T)$. Then

$$P_1 f = \sum_{i=1}^{3} c_i \phi_i$$

with $c_i = \int_T f\phi_i$. Therefore,

$$\frac{\partial P_1 f}{\partial x} = \frac{2\sqrt{6}}{h^{\frac{3}{2}}} \int_T f(\phi_2 - \phi_3) = \frac{24}{h^2} \int_T f(x, y)\Big(y + \frac{2x}{h} - 1\Big)\, dxdy.$$

Now observe that, for any $y \in (0, 1)$,

$$\int_0^{h(1-y)} \Big(y + \frac{2x}{h} - 1\Big)\, dx = 0$$

and so, denoting $\bar{f}(y) = \frac{1}{h(1-y)} \int_0^{h(1-y)} f(x, y)\ dx$, we obtain

$$\frac{\partial P_1 f}{\partial x} = \frac{24}{h^2} \int_0^1 \int_0^{h(1-y)} (f(x, y) - \bar{f}(y))\Big(y + \frac{2x}{h} - 1\Big)\, dxdy.$$

But using the one dimensional Poincaré inequality we have

$$\int_0^{h(1-y)} |f(x, y) - \bar{f}(y))|\, dx \le \frac{h}{2} \int_0^{h(1-y)} \Big|\frac{\partial f}{\partial x}(x, y)\Big|\, dx$$

and, since $\big|y + \frac{2x}{h} - 1\big| \le 1$, it follows that

$$\Big|\frac{\partial P_1 f}{\partial x}\Big| \le \frac{12}{h} \int_0^1 \int_0^{h(1-y)} \Big|\frac{\partial f}{\partial x}(x, y)\Big|\, dxdy.$$

Therefore

$$\Big|\frac{\partial P_1 f}{\partial x}\Big| \le \frac{12}{h} \Big\|\frac{\partial f}{\partial x}\Big\|_{L^1(T)} \le \frac{12}{h} |T|^{\frac{1}{2}} \Big\|\frac{\partial f}{\partial x}\Big\|_{L^2(T)}$$

and consequently

$$\Big\|\frac{\partial P_1 f}{\partial x}\Big\|_{L^2(T)} \le 6 \Big\|\frac{\partial f}{\partial x}\Big\|_{L^2(T)}.$$

Clearly, the same arguments can be applied to bound the derivative with respect to $y$. □

**Theorem 5.4.** *There exists a constant $C$ depending only on the maximum angle of $T$ such that*

$$\|\boldsymbol{u} - RT_1\boldsymbol{u}\|_{L^2(T)} \le Ch_T^2 \|D^2\boldsymbol{u}\|_{L^2(T)}.$$

*Proof.* From the definition of $RT_1\boldsymbol{u}$ we know that, for $i = 1, 2, 3$, $(\boldsymbol{u} - RT_1\boldsymbol{u}) \cdot \nu_i$ satisfies the hypotheses of Lemma 5.2 and then

$$\|(\boldsymbol{u} - RT_1\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \le Ch_T^2 \|D^2(\boldsymbol{u} - RT_1\boldsymbol{u})\|_{L^2(T)}.$$

So, in order to estimate the component of $\boldsymbol{u} - RT_1\boldsymbol{u}$ in the direction $\nu_i$, we need to bound the second derivatives of $RT_1\boldsymbol{u}$ in terms of $D^2\boldsymbol{u}$.

But an easy computation shows that, for any $\boldsymbol{v} \in \mathcal{RT}_1(T)$,

$$\frac{\partial^2 \boldsymbol{v}}{\partial x^2} = \frac{2}{3}\Big(\frac{\partial(\operatorname{div} \boldsymbol{v})}{\partial x}, 0\Big), \quad \frac{\partial^2 \boldsymbol{v}}{\partial y^2} = \frac{2}{3}\Big(0, \frac{\partial(\operatorname{div} \boldsymbol{v})}{\partial y}\Big)$$

and

$$\frac{\partial^2 \boldsymbol{v}}{\partial x \partial y} = \frac{1}{3}\Big(\frac{\partial(\operatorname{div} \boldsymbol{v})}{\partial y}, \frac{\partial(\operatorname{div} \boldsymbol{v})}{\partial x}\Big).$$

Therefore we have

$$\|(\boldsymbol{u} - RT_1\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \le Ch_T^2\big\{\|D^2\boldsymbol{u}\|_{L^2(T)} + \|\nabla(\operatorname{div}\ RT_1\boldsymbol{u})\|_{L^2(T)}\big\}. \tag{5.5}$$

Now from (5.1) we know that

$$\nabla(\operatorname{div} RT_1\boldsymbol{u}) = \nabla(P_1 \operatorname{div} \boldsymbol{u}),$$

hence applying Lemma 5.3 yields

$$\|\nabla(\operatorname{div} RT_1\boldsymbol{u})\|_{L^2(T)} \le C\|\nabla(\operatorname{div} \boldsymbol{u})\|_{L^2(T)}$$

and using this inequality in (5.5) we obtain the estimates for the normal components of $(\boldsymbol{u} - RT_1\boldsymbol{u})$. Then, to conclude the proof of the theorem, we proceed as in the case of $RT_0$. □

**5.2. The three-dimensional case.** As in the case of the Lagrange interpolation, the 3D case presents some important differences with the 2D one. We recall that the definition of $\mathcal{RT}_k$ can be extended straightforwardly to the 3D case. Indeed, for $T$ a tetrahedron we have

$$\mathcal{RT}_k(T) = \mathcal{P}_k^3(T) \oplus (x, y, z)\mathcal{P}_k(T).$$

The *maximum angle condition* can be generalized in different ways. The first one, introduced in [2], is the *regular vertex property*. We say that a tetrahedron satisfies this property with a constant $\bar{c} > 0$ if it has a vertex $\boldsymbol{p}_0$ such that $|\det M| \ge \bar{c} > 0$, where $M$ is the matrix which has $\boldsymbol{v}_i$, $i = 1, 2, 3$ as rows (where we are using the obvious generalization of the notation of the 2D case).

Under this hypothesis, Theorem 5.1 can be generalized almost straightforwardly. Indeed, the basic result given in Lemma 2.2 is valid now for functions with vanishing average on a face of $T$, and using this result we can prove, arguing as in the 2D case, that

$$\|(\boldsymbol{u} - RT_0\boldsymbol{u}) \cdot \nu_i\|_{L^2(T)} \le C \sum_{k=1}^{3} |\ell_k|\Big(\Big\|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}_k}\Big\|_{L^2(T)} + \|\operatorname{div} \boldsymbol{u}\|_{L^2(T)}|\nu_i \cdot \boldsymbol{v}_k|\Big).$$

As a consequence we obtain the following estimate.

**Theorem 5.5.** *Let $T$ be a tetrahedron satisfying the regular vertex property with a constant $\bar{c} > 0$. Then there exists a constant $C$ depending only on $\bar{c}$ such that*

$$\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)} \leq C \sum_{k=1}^{3} |\ell_k| \left( \left\| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}_k} \right\|_{L^2(T)} + \|\operatorname{div} \boldsymbol{u}\|_{L^2(T)} \right). \tag{5.6}$$

The other "natural" generalization of the 2D maximum angle condition is the condition introduced by Krízek [22]. We say that a family of tetrahedra satisfies the *maximum angle condition* with a constant $\overline{\psi} < \pi$ if the angles inside the faces and the angles between faces are bounded above by $\overline{\psi}$.

It is easy to see that in the 2D case the regular vertex property is equivalent to the maximum angle condition. However, the situation is different in the 3D case. In fact, the family in Figure 5, with arbitrary lengths $h_1, h_2, h_3$, satisfies uniformly the maximum angle condition but not the regular vertex property (take for example $h_1 = h_3 = h^2$, and $h_2 = h$). On the other hand, the regular vertex property implies the maximum angle condition (see [2]). A natural question is whether or not error

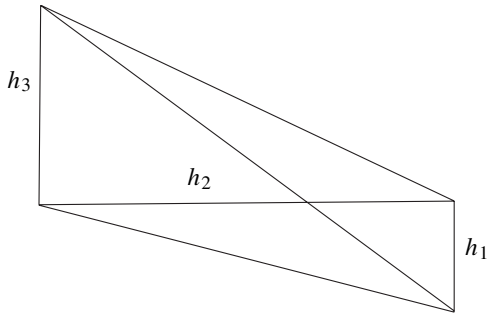



Figure 5

estimates for the $RT_0$ interpolation hold under the maximum angle condition. The answer is positive. In [2] the following result was proved.

**Theorem 5.6.** *If $T$ is a tetrahedron satisfying the maximum angle condition with a constant $\overline{\psi}$. Then there exists a constant $C$ depending only on $\overline{\psi}$ such that*

$$\|\boldsymbol{u} - RT_0\boldsymbol{u}\|_{L^2(T)} \leq Ch_T \|D\boldsymbol{u}\|_{L^2(T)}. \tag{5.7}$$

Again the basic tool to obtain this estimate is the generalization to 3D of Lemma 2.2. Indeed, consider the face mean average interpolator introduced in [15], namely, $\Pi\colon H^1(T) \to \mathcal{P}_1(T)$ given by

$$\int_S \Pi w = \int_S w$$

for any face $S$ of $T$.

**Lemma 5.7.** *The following error estimates hold with a constant $C$ independent of $T$:*

$$\|w - \Pi w\|_{L^2(T)} \le C \sum_{j=1}^{3} |\ell_j| \left\| \frac{\partial w}{\partial \boldsymbol{v}_j} \right\|_{L^2(T)} \tag{5.8}$$

$$\left\| \frac{\partial \Pi w}{\partial \xi} \right\|_{L^2(T)} \le \left\| \frac{\partial w}{\partial \xi} \right\|_{L^2(T)} \tag{5.9}$$

$$\left\| \frac{\partial (w - \Pi w)}{\partial \xi} \right\|_{L^2(T)} \le C \sum_{j=1}^{3} |\ell_j| \left\| \frac{\partial^2 w}{\partial \boldsymbol{v}_j \partial \xi} \right\|_{L^2(T)} \tag{5.10}$$

*where $\frac{\partial}{\partial \xi}$ is a derivative in any direction.*

*Proof.* Since $w - \Pi w$ has vanishing mean value on the faces of $T$, it follows from Lemma 2.2 that

$$\|w - \Pi w\|_{L^2(T)} \le C \sum_{j=1}^{3} |\ell_j| \left\| \frac{\partial (w - \Pi w)}{\partial \boldsymbol{v}_j} \right\|_{L^2(T)}. \tag{5.11}$$

Now, it follows from the definition of $\Pi$ that

$$\int_T \frac{\partial \Pi w}{\partial \xi} = \int_T \frac{\partial w}{\partial \xi}$$

or, in other words, the constant $\frac{\partial \Pi w}{\partial \xi}$ is the average on $T$ of $\frac{\partial w}{\partial \xi}$ and so (5.9) holds and (5.10) follows from Lemma 2.1. Finally, (5.8) is a consequence of (5.11) and (5.9). □

Now it is not difficult to check that, for any $\boldsymbol{u} \in H^1(T)^3$,

$$RT_0 \Pi \boldsymbol{u} = RT_0 \boldsymbol{u}$$

where $\Pi$ is the vector version of $\Pi$. Consequently,

$$\|\boldsymbol{u} - RT_0 \boldsymbol{u}\|_{L^2(T)} \le \|\boldsymbol{u} - \Pi \boldsymbol{u}\|_{L^2(T)} + \|\Pi \boldsymbol{u} - RT_0 \Pi \boldsymbol{u}\|_{L^2(T)}$$

and therefore, in view of (5.8), to prove (5.7) it is enough to prove the error estimate for $\boldsymbol{u} \in \mathcal{P}_1(K)^3$. In this way the problem is reduced to a finite dimensional one and the error estimate (5.7) can be proved under the maximum angle condition (see [2] for details).

## 6. The Stokes equations

The Stokes equations are given by

$$\begin{aligned} -\Delta \boldsymbol{u} + \nabla p &= f \quad \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u} &= 0 \quad \text{in } \Omega, \\ \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $\boldsymbol{u}$ is the velocity and $p$ the pressure of a fluid contained in $\Omega$.

This problem can be written in the form (1.1) with $V = H_0^1(\Omega)^n \times L_0^2(\Omega)$ where

$$L_0^2(\Omega) = \{f \in L^2(\Omega) \,:\, \textstyle\int_\Omega f = 0\},$$

$$B((\boldsymbol{u}, p), (\boldsymbol{v}, q)) = \sum_{i,j=1}^{n} \int_\Omega \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j} - \int_\Omega p \ \mathrm{div}\, \boldsymbol{v} - \int_\Omega q \ \mathrm{div}\, \boldsymbol{u}$$

and

$$F(\boldsymbol{v}, q) = \int_\Omega f\boldsymbol{v}.$$

Then to obtain a finite element approximation we need to use a space $W_h$ for the velocity and a space $Q_h$ for the pressure. Note that since in this case the form $B$ is symmetric, the two conditions (1.4) and (1.5) are exactly the same. From the classical theory for mixed finite elements of Brezzi [12] we know that to obtain (1.4) for the space $V_h = W_h \times Q_h$ it is enough to prove that there exists $\gamma > 0$, independent of $h$, such that

$$\inf_{q\in Q_h} \sup_{\boldsymbol{v}\in W_h} \frac{\int_\Omega q \ \mathrm{div}\, \boldsymbol{v}}{\|p\|_{L^2}\|\boldsymbol{v}\|_{H_0^1}} \geq \gamma. \tag{6.1}$$

Equivalently, for any $f \in Q_h$, there exists a solution $\boldsymbol{u} \in W_h$ of

$$\int_\Omega \mathrm{div}\, \boldsymbol{u} \cdot q = \int_\Omega fq \quad \text{for all } q \in Q_h, \tag{6.2}$$

$$\|\boldsymbol{u}\|_{H_0^1} \leq C\|f\|_{L^2} \tag{6.3}$$

with $C$ depending only on the domain $\Omega$.

A lot of work has been done to prove this inf-sup condition for different choices of spaces $W_h$ and $Q_h$. We refer for example to the books [13], [20]. However, most proofs require the regularity assumption (2.1) on the elements although it is not known whether it is essential or not.

One of the main tools to prove (6.1) is the so-called *Fortin operator* introduced in [19], which in the case of the Stokes equations is an operator $\Pi\colon H_0^1(\Omega)^n \to W_h$ such that

$$\int_\Omega q \ \mathrm{div}(\boldsymbol{v} - \Pi\boldsymbol{v}) = 0 \quad \text{for all } q \in Q_h$$

and

$$\|\Pi\boldsymbol{v}\|_{H_0^1} \leq C\|\boldsymbol{v}\|_{H_0^1} \tag{6.4}$$

with a constant $C$ independent of $h$.

Consider for example the non-conforming method of Crouzeix–Raviart, namely, $W_h$ are the $(\mathcal{P}_1)^n$ functions in each element which are also continuous at the midpoints of the edges or faces of the partition, and $Q_h$ are piecewise constant functions. Error estimates for anisotropic elements for this method have been proved in [2], [7].

The Fortin operator for this case is the edge (or face) mean average interpolator $\Pi$ defined in the previous section. In view of (5.9), estimate (6.4) holds with a constant independent of the geometry of the elements which can be taken to be one. However, this is a non-conforming method (because $W_h \not\subset H_0^1(\Omega)^2$) and therefore, to obtain error estimates, some consistency terms have to be bounded. This can be done by using the $RT_0$ interpolation analyzed in the previous section. In this way it is possible to obtain optimal error estimates for this method under the maximum angle condition (see [2]).

## References


[1] Acosta, G., Lagrange and average interpolation over 3D anisotropic elements. *J. Comp. Appl. Math.* **135** (2001), 91–109.

[2] Acosta, G., Durán, R. G., The maximum angle condition for mixed and non conforming elements: Application to the Stokes equations. *SIAM J. Numer. Anal.* **37** (2000), 18–36.

[3] Acosta, G., Durán, R. G., Error estimates for $\mathcal{Q}_1$ isoparametric elements satisfying a weak angle condition. *SIAM J. Numer. Anal.* **38** (2000), 1073–1088.

[4] Apel, T., *Anisotropic finite elements: Local estimates and applications*. Adv. Numer. Math., Teubner, Stuttgart 1999.

[5] Apel, T., Interpolation of non-smooth functions on anisotropic finite element meshes. *Math. Model. Numer. Anal.* **33** (1999), 1149–1185.

[6] Apel, T., Dobrowolski, M., Anisotropic interpolation with applications to the finite element method. *Computing* **47** (1992), 277–293.

[7] Apel, T., Nicaise, S., Schoeberl, J., Crouzeix-Raviart type finite elements on anisotropic meshes. *Numer. Math.* **89** (2001), 193–223.

[8] Arnold, D. N., Brezzi, F., Mixed and nonconforming finite element methods implementation, postprocessing and error estimates. *RAIRO, Modél. Math. Anal. Numér.* **19** (1985), 7–32.

[9] Babuška, I., Courant element: before and after. In *Finite element methods*, Lecture Notes Pure Appl. Math. 164, Dekker, New York 1994, 37–51.

[10] Babuška, I., Aziz, A. K., On the angle condition in the finite element method. *SIAM J. Numer. Anal.* **13** (1976), 214–226.

[11] Brenner, S. C., Scott, L. R., *The Mathematical Theory of Finite Element Methods*. Texts Appl. Math. 15, Springer-Verlag, New York 1994.

[12] Brezzi, F., On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *RAIRO Sér. Rouge* **8** (1974), 129–151.

[13] Brezzi, F., Fortin, M., *Mixed and Hybrid Finite Element Methods*. Springer Ser. Comput. Math. 15, Springer-Verlag, New York 1991.

[14] Ciarlet, P. G., *The Finite Element Method for Elliptic Problems*. Stud. Math. Appl. 4, North Holland, Amsterdam 1978.

[15] Crouzeix, M., Raviart, P. A., Conforming and non-conforming finite element methods for solving the stationary Stokes equations. *RAIRO Anal. Numér.* **7** (1973), 33–76.

[16] Durán, R. G., Error estimates for 3-d narrow finite elements. *Math. Comp.* **68** (1999), 187–199.

[17] Durán, R. G., Lombardi, A. L., Error estimates on anisotropic $\mathcal{Q}_1$ elements for functions in weighted Sobolev spaces. *Math. Comp.* **74** (2005), 1679–1706.

[18] Durán, R. G., Lombardi, A. L., Finite element approximation of convection diffusion problems using graded meshes. *Appl. Numer. Math.*, to appear.

[19] Fortin, M., An analysis of the convergence of mixed finite element methods. *RAIRO Anal. Numér.* **11** (1977), 341–354.

[20] Girault, V., Raviart, P. A., *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*. Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin 1986

[21] Jamet, P., Estimations d'erreur pour des éléments finis droits presque dégénérés. *RAIRO Anal. Numér.* **10** (1976), 46–71.

[22] Krízek, M., On the maximum angle condition for linear tetrahedral elements. *SIAM J. Numer. Anal.* **29** (1992), 513–520.

[23] Marini, L. D., An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method. *SIAM J. Numer. Anal.* **22** (1985), 493–496.

[24] Raviart, P. A., Thomas, J. M., A mixed finite element method for second order elliptic problems. In *Mathematical aspects of the Finite Element Method* (ed. by I. Galligani and E. Magenes), Lecture Notes in Math. 606, Springer-Verlag, Berlin 1977, 292–315.

[25] Roos, H. G., Stynes, M., Tobiska, L., *Numerical Methods for Singularly Perturbed Differential Equations*. Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin 1996.

[26] Shenk, N. A., Uniform error estimates for certain narrow Lagrange finite elements. *Math. Comp.* **63** (1994), 105–119.

Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina

E-mail: rduran@dm.uba.ar

# Linear subdivision schemes for the refinement of geometric objects

Nira Dyn

**Abstract.** Subdivision schemes are efficient computational methods for the design, representation and approximation of surfaces of arbitrary topology in $\mathbb{R}^3$. Subdivision schemes generate curves/surfaces from discrete data by repeated refinements. This paper reviews some of the theory of linear stationary subdivision schemes and their applications in geometric modelling. The first part is concerned with "classical" schemes refining control points. The second part reviews linear subdivision schemes refining other objects, such as vectors of Hermite-type data, compact sets in $\mathbb{R}^n$ and nets of curves in $\mathbb{R}^3$. Examples of various schemes are presented.

**Mathematics Subject Classification (2000).** Primary 65D07, 65D17, 65D18; Secondary 41A15, 68U07.

**Keywords.** Subdivision schemes, geometric modelling, curves, surfaces, refinements, smoothness, arbitrary topology, nets of points, nets of curves, compact sets, approximation order.

## 1. Introduction

Subdivision schemes in geometric applications are efficient tools for the generation of curves/surfaces from discrete data, by repeated refinements.

The first subdivision schemes where devised by de Rahm [54] for the generation of functions with a first derivative everywhere and a second derivative nowhere.

In geometric modelling the first schemes were proposed for easy and quick rendering of $B$-spline curves. A $B$-spline curve has the form

$$C(t) = \sum_i P_i B_m(t - i) \tag{1}$$

with $\{P_i\}$ points in $\mathbb{R}^d$ ($d = 2$ or 3) termed control points, and $B_m$ a $B$-spline of degree $m$ with integer knots, namely $B_m|_{[i,i+1]}$ is a polynomial of degree $m$, $B_m \in C^{m-1}(\mathbb{R})$, supp $B_m = [0, m+1]$. Equation (1) is a parametric representation of a $B$-spline curve. By using the refinement equation satisfied by a $B$-spline,

$$B_m(x) = \sum_{i=0}^{m+1} a_i^{[m]} B_m(2x - i), \quad a_i^{[m]} = 2^{-m}\binom{m+1}{i}, \qquad i = 0, \ldots, m+1. \tag{2}$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$C(t)$ in (1) has the parametric representations

$$C(t) = \sum_i P_i^0 B_m(t-i) = \sum_i P_i^1 B_m(2t-i) = \cdots \\ = \sum_i P_i^k B_m(2^k t - i) = \cdots, \tag{3}$$

where

$$P_i^{\ell+1} = \sum_j a_{i-2j}^{[m]} P_j^\ell, \quad \ell = 0, 1, 2, \ldots, \tag{4}$$

with the convention $a_i^{[m]} = 0$, $i \notin \{0, 1, \ldots, m+1\}$.

As is demonstrated in §2.3, the differences $\{P_i^k - P_{i-1}^k\}$ tend to zero as $k$ increases, and since $B_m \geq 0$ and $\sum_i B_m(t-i) \equiv 1$ [6], the polygonal line through the control points $\{P_i^k\}$ is close to $C(t)$ for $k$ large enough, and can be easily rendered.

The relation (4) encompasses the refinement rule for $B$-spline curves. The first scheme of this type was devised by Chaikin [10] for quadratic $B$-spline curves, and the schemes for general $B$-spline curves were investigated in [14]. All other subdivision schemes can be regarded as a generalization of the spline case.

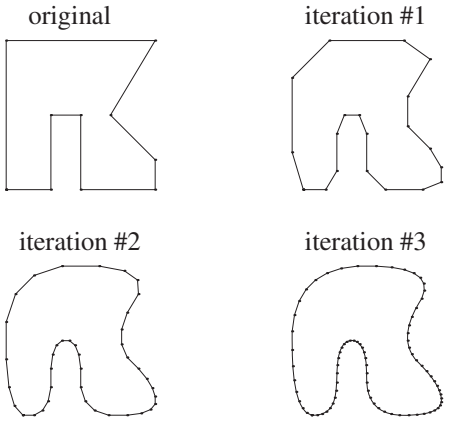



Figure 1. Refinements of a polygon with Chaikin scheme.

In this paper we first review the "classical" subdivision schemes for the refinement of control points. The schemes for the generation of curves are direct generalizations of (4), in the sense that the coefficients, defined in (2), are replaced by other sets of coefficients. The "classical" schemes, and in particular those generating surfaces, are used extensively in Computer Graphics. In §2 we discuss the construction of such schemes, their approximation properties, tools for the analysis of their convergence and smoothness, and their application to the generation of surfaces from general nets of points in $\mathbb{R}^3$. Examples of important schemes are presented.

Subdivision schemes for the refinement of objects other than control points are reviewed in §3. These schemes include subdivision schemes refining vectors, in particular, vectors consisting of values of a function and its consecutive derivatives, schemes refining compact sets in $\mathbb{R}^n$ and a scheme refining nets of curves.

All the schemes reviewed in this paper are linear. Recently, various non-linear schemes were devised and analyzed (see, e.g., [26] and references therein). It seems that this is one of the future directions in the study of subdivision schemes. Applications of "classical" schemes to the numerical solution of special types of PDEs is another direction. (See, e.g., [11]).

New "classical" schemes are still being devised for particular applications. For example, adaptive refinements can be accomplished straightforwardly by refining according to topological rules different from the "classical" ones, therefore, corresponding linear schemes had to be devised (see, e.g. [48] and [59]).

## 2. Stationary linear schemes for the refinement of control points

A subdivision scheme $S_{\boldsymbol{a}}$ for the refinement of control points is defined by a finite set of coefficients called *mask* $\boldsymbol{a} = \{a_i \in \mathbb{R} : i \in \sigma(\boldsymbol{a}) \subset \mathbb{Z}^s\}$. Here $\sigma(\boldsymbol{a})$ denotes the finite support of the mask, $s = 1$ corresponds to curves and $s = 2$ to surfaces. The refinement rule is

$$P_\alpha^{k+1} = \sum_{\beta \in \mathbb{Z}^s} a_{\alpha - 2\beta} P_\beta^k, \quad \alpha \in \mathbb{Z}^s. \tag{5}$$

**Remark.** In most of the paper we consider schemes defined on $\mathbb{Z}^s$, although, in geometric applications the schemes operate on finite sets of data. Due to the finite support of the mask, our considerations apply directly to closed curves/surfaces, and also to "open" ones, except in a finite zone near the boundary.

In the case $s = 1$, a subdivision scheme is termed *uniformly convergent* (or convergent for geometric applications) if the sequence $\{\mathcal{P}^k(t)\}$ of polygonal lines through the control points at refinement levels $k = 0, 1, 2, \ldots$ (with parametric representation as the piecewise linear interpolants to the data $\{(i2^{-k}, P_i^k) : i \in \mathbb{Z}\}$, $k = 0, 1, 2, \ldots$), converges uniformly in bounded intervals. In the case $s = 2$, we require the uniform convergence of the sequence of piecewise bi-linear interpolants to the data $\{(\alpha 2^{-k}, P_\alpha^k) : \alpha \in \mathbb{Z}^2\}$ on bounded squares [9], [33], [24].

The convergence of a scheme $S_{\boldsymbol{a}}$ implies the existence of a *basic-limit-function* $\phi_{\boldsymbol{a}}$, being the limit obtained from the initial data, $f_i^0 = 0$ everywhere on $\mathbb{Z}^s$ except $f_0^0 = 1$.

It follows from the linearity and uniformity of (5) that the limit obtained from any set of initial control points $\boldsymbol{P}^0 = \{P_\alpha^0 \in \mathbb{R}^d : \alpha \in \mathbb{Z}^s\}$, $S_{\boldsymbol{a}}^\infty \boldsymbol{P}^0$, can be written in terms of integer translates of $\phi_{\boldsymbol{a}}$, as

$$S_{\boldsymbol{a}}^\infty \boldsymbol{P}^0(x) = \sum_{\alpha \in \mathbb{Z}^s} P_\alpha^0 \phi_{\boldsymbol{a}}(x - \alpha), \quad x \in \mathbb{R}^s. \tag{6}$$

For $s = 1$ and $d = 2$ or $d = 3$, (6) is a parametric representation of a curve in $\mathbb{R}^d$, while for $s = 2$ and $d = 3$, (6) is a parametric representation of a surface in $\mathbb{R}^3$. Also, by the linearity, uniformity and stationarity of the refinement (5), $\phi_{\boldsymbol{a}}$ satisfies the *refinement equation* (*two-scale relation*)

$$\phi_{\boldsymbol{a}}(x) = \sum_{\alpha \in \mathbb{Z}^s} a_\alpha \phi_{\boldsymbol{a}}(2x - \alpha), \tag{7}$$

analogous to the refinement equation (2) for $B$-splines.

It follows from (5) or from (7) that $\operatorname{supp}(\phi_{\boldsymbol{a}})$ is contained in the convex hull of $\sigma(\boldsymbol{a})$ [9], as is the case for the $B$-spline schemes.

The choice of the mask in the design of good schemes is partly heuristic and partly aims at obtaining specific properties of the scheme as convergence, smoothness, locality, interpolation, shape preservation, and approximation order.

For the case $s = 1$, the topology of $\mathbb{Z}$ is sufficient to describe an ordered set of control points for curve design. For the case $s = 2$, the topology of $\mathbb{Z}^2$, where the point $(i, j)$ is connected to the four points $(i \pm 1, j)$, $(i, j \pm 1)$, is sufficient to describe a set of control points in $\mathbb{R}^3$, connected each to four neighboring points and constituting a *quad-mesh*. The above connectivity of $\mathbb{Z}^2$, with the additional connections of the point $(i, j)$ to the points $(i + 1, j + 1)$, $(i - 1, j - 1)$, forms the *three-direction mesh* which is sufficient to describe a regular triangulation (each vertex is connected to six neighboring vertices). These two types of topologies of $\mathbb{Z}^2$, are also relevant to general topologies of control points, since they are generated by most of the topological refinement rules. This is explained in §2.4.

**2.1. The main construction methods of schemes.** There are two main approaches to the construction of subdivision schemes. The first approach is by repeated averaging. In case $s = 1$, repeated averaging leads to $B$-spline schemes.

In this approach, the refinement rule (5) consists of several simple steps. The first is the trivial refinement

$$P_\alpha^{k+1,0} = P_{\left[\frac{\alpha}{2}\right]}^k, \quad \alpha \in \mathbb{Z}^s \tag{8}$$

with $\left[\frac{\alpha}{2}\right]$ the biggest integer smaller than or equal to $\frac{\alpha}{2}$ for $\alpha \in \mathbb{Z}$, and $\left(\left[\frac{\alpha_1}{2}\right], \left[\frac{\alpha_2}{2}\right]\right)$ for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{Z}^2$.

The trivial refinement is followed by a fixed number $m$ of repeated averaging

$$P_\alpha^{k+1,j} = \frac{P_\alpha^{k+1,j-1} + P_{\alpha - e_j}^{k+1,j-1}}{2}, \quad \alpha \in \mathbb{Z}^s, \ j = 1, \ldots, m,$$

where $\{e_1, \ldots, e_m\}$ are non-zero vectors in $\mathbb{Z}^s$ with components in $\{0, 1\}$.

The case $s = 1$ corresponds to the $B$-spline scheme of degree $m$, while in case $s = 2$ one gets the tensor-product $B$-spline schemes for the choice $e_1 = \cdots = e_r = (1, 0)$, $e_{r+1} = \cdots = e_m = (0, 1)$, $1 \le r < m$, and the three-direction box-spline schemes [7] for the choice $e_1 = \cdots = e_r = (1, 0)$, $e_{r+1} = \cdots = e_\rho = (0, 1)$,

$e_{\rho+1} = \cdots = e_m = (1, 1)$, $1 \le r < \rho < m$. One can get other box-spline schemes for more general choices of $e_1, \ldots, e_m$ [7].

The second construction of subdivision schemes is based on a local approximation operator $A$, approximating on $[0, 1]^s$. $A$ is defined in terms of samples of the approximated function in a set of points $\mathcal{A} \subset \mathbb{Z}^s$,

$$(Af)(x) = \sum_{\alpha \in \mathcal{A}} f(\alpha) w_\alpha(x), \quad x \in [0, 1]^s. \tag{9}$$

For geometrical applications, the set $\mathcal{A}$ contains the set $E_s$ of extreme points of $[0, 1]^s$, and is symmetric relative to $[0, 1]^s$. The operator $A$ has to be scale and shift invariant, so that (9) can be used in any refinement level and at any location. This leads to the choice of a polynomial approximation operator $A$.

The commonly derived refinement rule from (9) is

$$P^{k+1}_{2\alpha+\gamma} = \sum_{\beta \in \mathcal{A}} P^k_{\alpha+\beta} w_\beta \left(\frac{\gamma}{2}\right), \quad \gamma \in E_s. \tag{10}$$

Another possibility is

$$P^{k+1}_{2\alpha+\gamma} = \sum_{\beta \in \mathcal{A}} P^k_{\alpha+\beta} w_\beta \left(g + \frac{\gamma}{2}\right), \quad \gamma \in E_s,\ g = \{1/4\}^s. \tag{11}$$

In case $s = 1$, with $Af$ the interpolation polynomial based on the symmetric set of points relative to $[0, 1]$, $-N+1, \ldots, 0, 1, \ldots, N$, the resulting family of schemes obtained by (10) for $N = 1, 2, \ldots$ consists of the Dubuc–Deslaurier schemes [22]

$$P^{k+1}_{2i} = P^k_i, \quad P^{k+1}_{2i+1} = \sum_{\ell=-N+1}^{N} w_\ell \left(\frac{1}{2}\right) P^k_{i+\ell}, \quad w_i(x) = \prod_{\substack{j=-N+1 \\ j \ne i}}^{N} \frac{x - j}{i - j} \tag{12}$$

The schemes in (12) are interpolatory, since the set of control points after refinement contains the control points before refinement. These schemes are convergent, and the limit curves interpolate the initial control points [22]. Interpolatory schemes in general are discussed in [34].

Recently this construction was extended to non-interpolatory schemes [30], by using (11) instead of (10) with $w_i(x)$ defined in (12).

The refinement rules are

$$P^{k+1}_{2i} = \sum_{\ell=-N+1}^{N} w_\ell \left(\frac{1}{4}\right) P^k_{i+\ell}, \quad P^{k+1}_{2i+1} = \sum_{\ell=-N+1}^{N} w_\ell \left(\frac{3}{4}\right) P^k_{i+\ell}.$$

It is checked in [30] that, for $N \le 10$, the schemes are convergent with limit curves of higher smoothness than the limit curves of the corresponding Dubuc–Deslaurier schemes. Yet, there is no proof that this holds in general.

In fact, (11) can be further extended to

$$P^{k+1}_{2\alpha+\gamma} = \sum_{\beta\in\mathcal{A}} P^k_{\alpha+\beta} w_\beta\big(g + (1-2\mu)\gamma\big), \quad \gamma \in E_s,\ g = \{\mu\}^s$$

with $0 < \mu < \frac{1}{2}$.

This refinement was studied in [33], [40], [4], for $s = 1$ and $A$ a linear interpolation operator at the points $x = 0$, $x = 1$. For $\mu = \frac{1}{4}$, this is the Chaikin scheme for generating quadratic $B$-spline curves [10]. For $\mu \neq \frac{1}{4}$ it is a general *corner cutting* scheme.

**2.2. Approximation order of subdivision schemes.** A convergent subdivision scheme $S$, constructed by the second approach of §2.1 with refinement rule (10), has the property of *reproduction of polynomials*.

Let the operator $A$ map the set $f|_{\mathcal{A}}$ to a unique interpolation polynomial of total degree not exceeding $m$, interpolating the data $\{(x, f(x)) : x \in \mathcal{A}\}$. In the following, we denote by $\Pi_m(\mathbb{R}^s)$ the space of all $s$-variate polynomials of degree up to $m$. It is easy to verify that for $f \in \Pi_m(\mathbb{R}^s)$ and $\boldsymbol{f}^0 = \{f^0_\alpha = f(\alpha h) : \alpha \in \mathbb{Z}^s\}$, $h \in \mathbb{R}_+$, the refinement (10) generates data on $f$, namely $\boldsymbol{f}^k = S^k \boldsymbol{f}^0 = \{f^k_\alpha = f(2^{-k}\alpha h) : \alpha \in \mathbb{Z}^s\}$, and therefore $S^\infty \boldsymbol{f}^0 = f$, and the subdivision scheme reproduces polynomials in $\Pi_m(\mathbb{R}^s)$.

In case of the refinement rule (11), arguments as in [30] lead to $(S^\infty \boldsymbol{f}^0)(x) = f(x + 2hg)$, with $g$ as in (11). This property of the scheme $S$ is termed reproduction *with a fixed shift* of polynomials in $\Pi_m(\mathbb{R}^s)$.

The reproduction of polynomials in $\Pi_m(\mathbb{R}^s)$ (with or without a shift), the representation of $S^\infty \boldsymbol{f}^0$ in terms of the compactly supported basic limit function $\phi$ of $S$,

$$S^\infty \boldsymbol{f}^0(x) = \sum_{\alpha\in\mathbb{Z}^s} f^0_\alpha \phi(x - \alpha), \tag{13}$$

and classical quasi-interpolation arguments [5], lead to the error estimate

$$\sup_{x\in\Omega} \big|(S^\infty \boldsymbol{f}^0)(x) - f(x)\big| \le C h^{m+1}. \tag{14}$$

In (14) $\boldsymbol{f}^0 = \{f^0_\alpha = f(\alpha h) : \alpha \in \mathbb{Z}^s\}$ for the refinement rule (10), while, for the refinement rule (11), $\boldsymbol{f}^0 = \{f^0_\alpha = f(\alpha h - 2gh) : \alpha \in \mathbb{Z}^s\}$, where $f$ is a smooth enough function, $\Omega$ is a bounded domain in $\mathbb{R}^s$, and the constant $C$ may depend on $S$, $f$, $\Omega$ but not on $h$. A subdivision scheme satisfying (14) is said to have *approximation order* $m + 1$.

Subdivision schemes constructed by repeated averaging reproduce constant functions and hence have approximation order 1. If the repeated averaging is done in a symmetric way relative to $[0, 1]^s$, then the resulting scheme reproduces also linear polynomials, and the scheme has approximation order 2. For example, this property is shared by all the symmetric $B$-spline schemes of odd degrees. The mask of the

scheme generating $B$-spline curves, based on the symmetric $B$-spline of degree $2\ell+1$ is

$$\widetilde{a}_i^{[2\ell+1]} = \frac{1}{2^{2\ell+1}}\binom{2\ell+2}{\ell+1+i}, \quad i = -\ell-1, \ldots, 0, \ldots, \ell+1.$$

The repeated averaging for such a symmetric mask takes the symmetric form

$$P_{2i}^{k+1,0} = P_i^k, \quad P_{2i+1}^{k+1,0} = \frac{1}{2}(P_i^k + P_{i+1}^k), \quad i \in \mathbb{Z},$$
$$P_i^{k+1,j} = \frac{1}{4}\big(P_{i-1}^{k+1,j-1} + 2P_i^{k+1,j-1} + P_{i+1}^{k+1,j-1}\big), \quad i \in \mathbb{Z},\ j = 1, \ldots, \ell,$$
$$P_i^{k+1} = P_i^{k+1,\ell}, \quad i \in \mathbb{Z}.$$

**2.3. Convergence and smoothness analysis.** Given the coefficients of the mask of a scheme, one would like to be able to determine if the scheme is convergent, and what is the smoothness of the resulting basic limit function (which is the generic smoothness of the limits generated by the scheme in view of (13)). Such analysis tools are essential for the design of new schemes.

We present one method for convergence analysis of the two cases $s = 1, 2$. The method for smoothness analysis in case $s = 1$ is simpler and is given in full. Its extension to $s = 2$ is omitted, but some special cases are discussed. There are other methods for convergence and smoothness analysis, see, e.g., [18], [19], [20], [42], [45].

An important tool in the analysis of convergence, presented here, is the symbol of a scheme $S_{\boldsymbol{a}}$ with the mask $\boldsymbol{a} = \{a_\alpha : \alpha \in \sigma(\boldsymbol{a})\}$,

$$a(z) = \sum_{\alpha\in\sigma(\boldsymbol{a})} a_\alpha z^\alpha. \tag{15}$$

A first step towards the convergence analysis is the derivation of the necessary condition for uniform convergence,

$$\sum_{\beta\in\mathbb{Z}^s} a_{\alpha-2\beta} = 1, \quad \alpha \in E_s, \tag{16}$$

derived easily from the refinement rule

$$f_\alpha^{k+1} = \sum_{\beta\in\mathbb{Z}^s} a_{\alpha-2\beta} f_\beta^k, \quad \alpha \in \mathbb{Z}^s.$$

with $\boldsymbol{f}^k = \{f_\alpha^k \in \mathbb{R} : \alpha \in \mathbb{Z}^s\}$. The necessary condition (16) implies that we have to consider symbols satisfying

$$a(1) = 2, \quad a(-1) = 0 \quad \text{if } s = 1, \tag{17}$$

or

$$a(1,1) = 4, \quad a(-1,1) = a(1,-1) = a(-1,-1) = 0 \quad \text{if } s = 2. \tag{18}$$

In case $s = 1$, condition (17) is equivalent to

$$a(z) = (1 + z)q(z) \quad \text{with} \quad q(1) = 1. \tag{19}$$

The scheme with symbol $q(z)$, $S_{\boldsymbol{q}}$, satisfies $S_{\boldsymbol{q}}\Delta = \Delta S_{\boldsymbol{a}}$ (see, e.g. [24]), where $\Delta$ is the difference operator

$$\Delta \boldsymbol{f} = \left\{(\Delta \boldsymbol{f})_i = f_i - f_{i-1} : i \in \mathbb{Z}\right\}. \tag{20}$$

A necessary and sufficient condition for the convergence of $S_{\boldsymbol{a}}$ is the contractivity of the scheme $S_{\boldsymbol{q}}$, namely $S_{\boldsymbol{a}}$ is convergent if and only if $S_{\boldsymbol{q}}^{\infty}\boldsymbol{f}^0 = 0$ for any $\boldsymbol{f}^0$ [33]. The contractivity of $S_{\boldsymbol{q}}$ is equivalent to the existence of a positive integer $L$, such that $\|S_{\boldsymbol{q}}^L\|_\infty < 1$. This condition can be checked for a given $L$ by algebraic operations on the symbol $q(z)$ (see, e.g., [24], [25]).

For practical geometrical reasons, only small values of $L$ have to be considered, since a small value of $L$ guarantees "visual convergence" of $\{\mathcal{F}^k(t)\}$ to $S_{\boldsymbol{a}}^{\infty}\boldsymbol{P}^0$, already for small $k$, as the distances between consecutive control points contract to zero fast. A good scheme corresponds to $L = 1$ as the $B$-spline schemes, or to $L = 2$ as many of the schemes constructed by the second method in §2.1 (see the following examples).

For $s = 2$, the necessary condition (18) guarantees the existence of two decompositions of the form

$$(1 - z_i)a(z) = q_{i1}(z)(1 - z_1^2) + q_{i2}(z)(1 - z_2^2), \quad i = 1, 2, \tag{21}$$

where $z = (z_1, z_2)$. The above two decompositions extend to $s = 2$ the factorization (19) written as $(1 - z)a(z) = (1 - z^2)q(z)$. The decompositions (21) guarantee the existence of a matrix subdivision scheme $S_Q$, with a $2 \times 2$ matrix symbol $Q(z) = \{q_{ij}(z)\}_{i,j=1}^2$, satisfying $S_{\boldsymbol{Q}}(\Delta_1, \Delta_2)^T = (\Delta_1, \Delta_2)^T S_{\boldsymbol{a}}$. Here $(\Delta_1, \Delta_2)^T$ is the vector difference operator, extending (20) to $s = 2$,

$$(\Delta_1, \Delta_2)^T \boldsymbol{f} = \left\{((\Delta_1, \Delta_2)^T \boldsymbol{f})_\alpha = (f_\alpha - f_{\alpha-(1,0)}, f_\alpha - f_{\alpha-(0,1)})^T : \alpha \in \mathbb{Z}^2\right\}.$$

A sufficient condition for the convergence of $S_{\boldsymbol{a}}$ is the contractivity of $S_{\boldsymbol{Q}}$, which can be checked by algebraic operations on the symbol $Q(z)$ [9], [24], [44].

Since many of the schemes have symmetries relative to $\mathbb{Z}^2$, their symbols are factorizable and have the form $a(z) = (1 + z_1)(1 + z_2)q(z)$. As a simple extension of the case $s = 1$, we get that $S_{\boldsymbol{a}}$ is convergent if the two schemes with symbols $(1 + z_1)q(z)$, $(1 + z_2)q(z)$ are contractive. If $a(z)$ is symmetric in the sense that $q(z_1, z_2) = q(z_2, z_1)$, then it is sufficient to check the contractivity of $(1 + z_1)q(z)$ (see, e.g., [25]).

The smoothness analysis in the case $s = 1$, relies on the result that if the symbol of a scheme has a factorization

$$a(z) = \left(\frac{1 + z}{2}\right)^{\nu} b(z), \tag{22}$$

such that the scheme $S_{\boldsymbol{b}}$ is convergent, then $S_{\boldsymbol{a}}$ is convergent and its limit functions are related to those $S_{\boldsymbol{b}}$ by

$$D^{\nu}(S_{\boldsymbol{a}}^{\infty}\boldsymbol{f}^{0}) = S_{\boldsymbol{b}}^{\infty}\Delta^{\nu}\boldsymbol{f}^{0}, \tag{23}$$

with $D$ the differentiation operator [33], [24]. Thus, each factor $(1+z)/2$ multiplying a symbol of a convergent scheme adds one order of smoothness. This factor is termed a *smoothing factor*.

The relation between (22) and (23) is a particular instance of the "algebra of symbols" [35]. If $a(z)$, $b(z)$ are two symbols of converging schemes, then $S_c$ with the symbol $c(z) = \frac{1}{2^s}a(z)b(z)$ is convergent, and

$$\phi_{\boldsymbol{c}} = \phi_{\boldsymbol{a}} * \phi_{\boldsymbol{b}}. \tag{24}$$

**Example ($B$-spline schemes).** The smoothness of the limit functions generated by the $m$-th degree $B$-spline scheme, having the symbol $a^{[m]}(z) = 2\left(\frac{1+z}{2}\right)^{m+1}$, can be concluded easily. The factor $b(z) = \frac{(1+z)}{2}^2$ corresponds to $S_{\boldsymbol{b}}$ generating a piecewise linear interpolant to the initial data $\{(i,\, f_i^0)\}$, which is continuous, and the factors $\left(\frac{1+z}{2}\right)^{m-1}$ add smoothness, so that $S_{\boldsymbol{a}^{[m]}}^{\infty} f^0 \in C^{m-1}$. Note that $a^{[m]}(z)$ consists of smoothing factors only. In fact the $B$-spline schemes are optimal, in the sense that for a given support size of the mask, the limit functions generated by the corresponding $B$-spline scheme is of maximal smoothness.

**Example (the four-point scheme).** Here we present the most general univariate interpolatory scheme which is based on four points [31], and describe briefly its convergence and smoothness analysis.

The refinement rule is

$$f_{2i}^{k+1} = f_i^k, \quad f_{2i+1}^{k+1} = -w(f_{i-1}^k + f_{i+2}^k) + \left(\tfrac{1}{2} + w\right)(f_i^k + f_{i+1}^k),$$

with $w$ a parameter controlling the shape of the limit curves. The symbol of the scheme is

$$a_w(z) = \tfrac{1}{2z}(z+1)^2\left[1 - 2wz^{-2}(1-z)^2(z^2+1)\right]. \tag{25}$$

Note that for $w = 0$, $a_0(z)$ is the symbol of the two-point scheme generating the polygonal line through the initial control points, and that for $w = 1/16$ it coincides with the symbol of the Dubuc–Deslauriers scheme based on four points (reproducing cubic polynomials).

The range of $w$ for which $S_{\boldsymbol{a}_w}$ is convergent is the range for which $S_{\boldsymbol{b}_w}$ with symbol $b_w(z) = a_w(z)/(1+z)$ is contractive. The condition $\|S_{\boldsymbol{b}_w}\|_\infty < 1$ holds in the range $-3/8 < w < (-1+\sqrt{13})/8$, while the condition $\|S_{\boldsymbol{b}_w}^2\|_\infty < 1$ holds in the range $-1/4 < w < (-1+\sqrt{17})/8$. Thus $S_{\boldsymbol{a}_w}$ is convergent in the range $-3/8 < w < (-1+\sqrt{17})/8$. To find a range of $w$ where $S_{\boldsymbol{a}_w}$ generates $C^1$ limits,

the contractivity of $S_{\boldsymbol{c}_w}$ with $c_w(z) = 2a_w(z)/(1+z)^2$ has to be investigated. It is easy to check that $\|S_{\boldsymbol{c}_w}\|_\infty \geq 1$, but that $\|S^2_{\boldsymbol{c}_w}\|_\infty < 1$ for $0 < w < (\sqrt{5}-1)/8$.

The limit of $S_{\boldsymbol{a}_w}$ is not $C^2$ even for $w = 1/16$, although for $w = 1/16$ the symbol is divisible by $(1+z)^3$ (see, e.g., [31]). It is shown in [20] by other methods, that the basic limit function for $w = 1/16$, restricted to its support, has a second derivative only at the non-dyadic points.

For the case $s = 2$, the idea of smoothing factors generalizes straightforwardly. Two smoothing factors in two linearly independent directions in $\mathbb{Z}^2$ are sufficient for increasing the smoothness. A smoothing factor in direction $(u, v) \in \mathbb{Z}^2$ is $\frac{1}{2}(1+z_1^u z_2^v)$. Specializing to the coordinate directions in $\mathbb{Z}^2$, $(1, 0)$ and $(0, 1)$, we get for a symbol $a(z) = (1+z_1)^m(1+z_2)^m b(z)$, such that $S_{\boldsymbol{b}}$ is convergent, that

$$\partial_{i,j} S^\infty_{\boldsymbol{a}} f^0 = S^\infty_{\boldsymbol{a}_{i,j}} \Delta_1^i \Delta_2^j f^0, \quad i, j = 0, \ldots, m, \tag{25}$$

with

$$a_{i,j}(z) = \frac{2^{i+j} a(z)}{(1+z_1)^i (1+z_2)^j}, \quad i, j = 0, \ldots, m, \tag{26}$$

and with $\partial_{ij}$ the $(i+j)$-th partial derivative of orders $i, j$ in directions $(1, 0)$ and $(0, 1)$ respectively.

For a symbol with the symmetry of the three direction mesh

$$a(z) = (1+z_1)^m (1+z_2)^m (1+z_1 z_2)^m b(z), \tag{27}$$

such that $S_{\boldsymbol{b}}$ is convergent, we get

$$\partial_{i,j,\ell} S^\infty_{\boldsymbol{a}} f^0 = S^\infty_{\boldsymbol{a}_{i,j,\boldsymbol{\ell}}} \Delta_1^i \Delta_2^j (\Delta_1 + \Delta_2)^\ell f^0, \quad i, j, \ell = 0, \ldots, m, \tag{28}$$

with

$$a_{i,j,\ell}(z) = \frac{2^{i+j+\ell} a(z)}{(1+z_1)^i (1+z^2)^j (1+z_1 z_2)^\ell}, \quad i, j, \ell = 0, \ldots, m, \tag{29}$$

and with $\partial_{i,j,\ell}$ the $(i+j+\ell)$-th partial derivative of orders $i, j, \ell$ in directions $(1, 0)$, $(0, 1)$, $(1, 1)$ respectively.

In particular $S_{\boldsymbol{a}}$ with the symbol $a(z) = (1+z_1)^2(1+z_2)^2 b(z)$ generates $C^1$ limit functions if the three schemes with the symbols

$$2(1+z_1)(1+z_2)b(z), \quad 2(1+z_1)^2 b(z), \quad 2(1+z_2)^2 b(z),$$

are contractive. Similarly for $a(z) = (1+z_1)(1+z_2)(1+z_1 z_2)b(z)$, $\phi_{\boldsymbol{a}} \in C^1$ if two of the three schemes with the symbols $2(1+z_1)b(z)$, $2(1+z_2)b(z)$, $2(1+z_1 z_2)b(z)$ are contractive.

The conditions for smoothness given above are only sufficient. Yet, in the case $s = 1$, there is a large class of convergent schemes for which the factorization in (22) is necessary for generating $C^\nu$ limit functions [39]. The schemes in this class are $L_\infty$-stable, namely, satisfy

$$\|S^\infty_a \boldsymbol{f}^0\| \geq C \|\boldsymbol{f}^0\|_\infty, \quad \boldsymbol{f}^0 \in \ell_\infty(\mathbb{Z}), \tag{31}$$

with constant $C$ dependent on $S_a$ but not on $\boldsymbol{f}^0$. All relevant schemes for geometric applications are $L_\infty$-stable, as the interpolatory schemes and the $B$-spline schemes.

This is not the case for $s = 2$. The symbol of a convergent $L_\infty$-stable scheme, generating smooth limit functions is not necessarily factorizable. Yet, many of the schemes in use have factorizable symbols.

**2.4. Subdivision schemes generating surfaces.** Schemes generating surfaces operate on control nets, and map a control net to a refined one.

A control net $N(V, E, F)$, consists of a set $V$ of points in $\mathbb{R}^3$, termed *vertices*, with two sets of topological relations between them $E$ and $F$, called *edges* and *faces* respectively (see, e.g., [49]). An edge denotes a pair of vertices. A face is a cyclic list of vertices where every pair of consecutive vertices constitutes an edge. The *valency* of a vertex is the number of edges that share it, the *valency* of a face is the number of vertices that belong to it. In Figure 2 we present a schematic net. We consider here

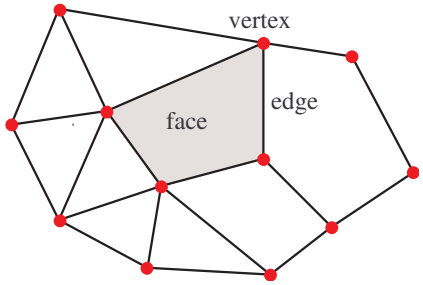



Figure 2. A schematic net.

only closed nets, namely nets in which each edge is shared by two faces.

**2.4.1. Topological refinement of nets.** There are several topological rules for refining a net $N(V, E, F)$. The most common one defines the new set of vertices, as

$$V' = \{u(v) : v \in V\} \cup \{u(e) : e \in E\} \cup \{u(f) : f \in F\} = V'_V \cup V'_E \cup V'_F. \quad (32)$$

Here $V'_V$ denotes all the new vertices, called *v-vertices*, corresponding to the vertices in $V$ (in an interpolatory scheme $V'_V = V$); $E'_V$ denotes all the new vertices, called *e-vertices*, corresponding to the edges in $E$, and $V'_F$ denotes all the new vertices, called *f-vertices*, corresponding to the faces in $F$. The rule for determining the location in $\mathbb{R}^3$ of $u(v)$, $u(e)$ and $u(f)$ is the refinement rule of the subdivision scheme. For example, a new vertex $u(e)$ is a certain linear combination of the vertices in $V$, weighted according to the topological relation between each $v \in V$ and $e$.

The topological relations $E'$, $F'$ in the refined net $N'(V', E', F')$ are independent of the subdivision scheme, but depend only on $E$ and $F$,

$$E' = \{(u(e), u(f)) : e \in f \in F\} \cup \{(u(e), u(v)) : v \in e \in E\} = E'_F \cup E'_E \quad (33)$$

and

$$F' = \{(u(v), u(e), u(f), u(\widetilde{e})) : v = e \cap \widetilde{e} \in f \in F\}. \quad (34)$$

Thus after one refinement step all faces have valency four and similarly all the vertices in the set $V'_E$. The valency of a vertex in $V'_F$ is the same as that of the "parent" face, and the valency of a vertex in $V'_V$ is the same as that of the "parent" vertex. From this observation we conclude that the nets obtained after two or more refinements have the topology of a quad-mesh (of $\mathbb{Z}^2$), except for a finite number of vertices with valency different from four (each equals the valency of an "ancestor" face or vertex in the initial net). The vertices with valency different from four are termed *irregular* (*extra-ordinary*) and a special local analysis of convergence and smoothness is required there [55]. Over the net, except in the vicinity of the irregular vertices, the analysis relative to $\mathbb{Z}^2$ is applicable.

For a net $N(V, E, F)$ with all faces of valency three, the topological refinement which is commonly used is such that the new vertices consist of $v$-vertices and $e$-vertices only, with the topological refinement

$$E' = E'_E \cup E'_V, \quad F' = F'_V \cup F'_F. \quad (35)$$

In (35), $E'_E$ is defined as in (33), and $E'_V = \{(u(e), u(\widetilde{e})) : e \cap \widetilde{e} \in V\}$. The new faces are of two types, $F'_V = \{(u(v), u(e), u(\widetilde{e})) : v \in e \cap \widetilde{e} \in V\}$, and $F'_F = \{(u(e_1), u(e_2), u(e_3)) : e_1, e_2, e_3 \in f \in F\}$. This refinement is presented in a schematic way in Figure 3. As can be observed from Figure 3, every face is replaced

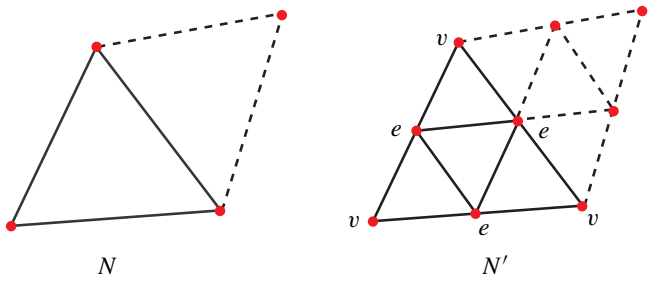



Figure 3. Schematic triangular topological refinement.

by four faces, one determined by the face itself, and three in $F'_V$, each consisting of three new vertices, one corresponding to one vertex of the face and two to the two edges of the face sharing that vertex.

Note that a face with valency three can be realized in $\mathbb{R}^3$ as a planar triangle, and therefore $N(V, E, F)$, with all faces of valency three, can be realized as a triangulation

of the set $V$. According to the topological refinement (35), the $e$-vertices have valency six, while a $v$-vertex has the same valency as that of its "parent" vertex in $V$. Thus, after two or more topological refinements, most of the vertices in the triangulations have valency six. Only a finite set of *irregular* (*extra-ordinary*) vertices have valencies different from six, "inherited" from those in the initial triangulation. Also, each irregular vertex is connected by edges only to regular vertices (of valency six).

Thus for a triangulation refined as above, the analysis of convergence and smoothness relative to the three-direction mesh applies, except in the vicinity of a finite number of isolated points, where a special local analysis is required [55].

While the analysis on regular meshes can handle any order of smoothness, the analysis at irregular vertices is limited to $C^1$ smoothness (see, [52], [46], and references therein). This limitation is the main reason why subdivision schemes are used mainly in computer graphics. In many industrial applications the designed surfaces have to be $C^2$ everywhere.

**2.4.2. Some popular schemes.** The first schemes devised for general nets were the bivariate tensor-product $B$-spline schemes of low degree, with special rules near irregular vertices [8], [23]. A bivariate tensor-product scheme of a univariate scheme with symbol $a(z)$, is a scheme with the symbol $a(z_1, z_2) = a(z_1)a(z_2)$.

The most commonly used scheme of that type for the topological refinement (32)–(34) is the Catmull–Clark scheme, which is an extension of the tensor-product cubic $B$-spline scheme [8]. The weights, up to normalization, of this scheme are given in Figure 4. The points designated by $o$ are the new $f$-vertices, and the weight of a vertex

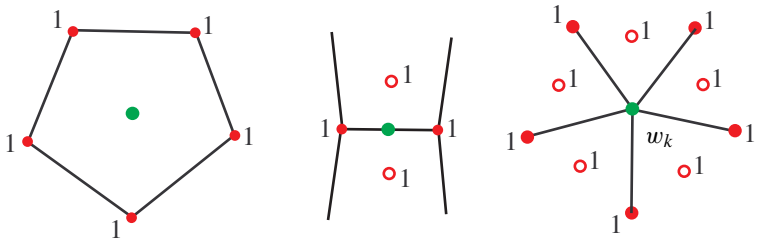



Figure 4. Weights for Catmull–Clark scheme: $f$-vertex (left), $e$-vertex (middle) and $v$-vertex (right).

of valency $k$, in the rule for its "son", is $w_k = k(k-2), k = 3, 4, \ldots$. Note that $w_4 = 8$, which is the weight in the tensor-produce cubic $B$-spline scheme. This scheme is easy to implement as can be inferred from Figure 4. Different choices of $w_k$ were considered in [2], [3] to improve the limit curvature at irregular vertices. Applications of the Catmull–Clark scheme are numerous. Here we refer to two important papers [21], [41].

In [47], the tensor-product four-point scheme is extended to an interpolatory scheme for general nets with the topological refinement (32)–(34).

For triangulations, the box-spline-based scheme of Loop [50] is very popular. Loop scheme is an extension of a box-spline scheme with the symbol

$$a(z_1, z_2) = \frac{1}{16}(1+z_1)^2(1+z_2)^2(1+z_1z_2)^2,$$

generating $C^2$ piecewise quartic box-spline surfaces on the three-direction mesh. The support of the mask of Loop scheme is small, and the refinement rule involves only neighboring vertices. In Figure 5 the weights for defining a new $e$-vertex and a new

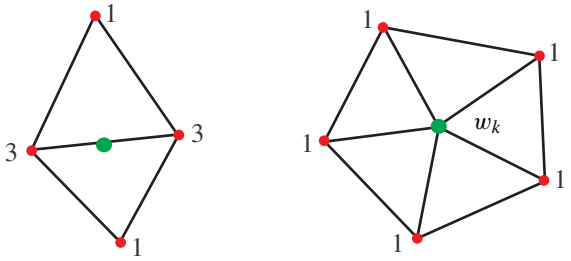



Figure 5. Weights for Loop scheme: $e$-vertex (left) and $v$-vertex (right).

$v$-vertex are given up to normalization. The weight $w_k$ of a vertex of valency $k$, involved in the rule for its "son", is

$$w_k = \frac{64k}{40 - \left(3 + 2\cos\frac{2\pi}{k}\right)}, \quad k = 3, 4, \ldots \tag{36}$$

Figure 6 depicts an initial triangulation of a head, and the triangulations after two refinements with Catmull–Clark scheme and with Loop scheme.

An interpolatory scheme for general closed triangulations with a shape parameter is the butterfly scheme [32]. The weights defining a new $e$-vertex are depicted in the left figure of Figure 7. Since the scheme is interpolatory, the new $v$-vertices coincide with the old vertices. The scheme generates $C^1$ surfaces if all vertices have valencies at least four and at most eight, depending on the value of $w$ [44], [57]. Modified weights for $e$-vertices, corresponding to edges having an irregular vertex of any valency above three are derived in [61] for $w = 1/16$. These weights are depicted in the right figure of Figure 7. The values $\{s_j\}$ are given by a formula depending on the valency $k$ of the irregular vertex,

$$s_j = \frac{1}{k}\left(\frac{1}{4} + \cos\frac{2\pi j}{k} + \frac{1}{2}\cos\frac{4\pi j}{k}\right), \quad j = 0, 1, \ldots, k-1,\ k > 3. \tag{37}$$

With the modified weights, the generated surfaces are $C^1$ for any valency greater than three, and are better looking in the vicinity of irregular vertices of valency between four and eight.

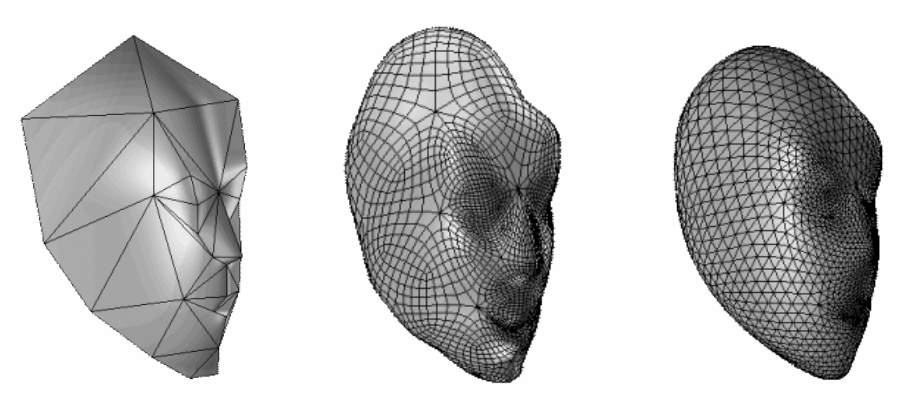


Figure 6. Head. Initial control net (left), after two refinements: with Catmull–Clark scheme (middle) and with Loop scheme (right).

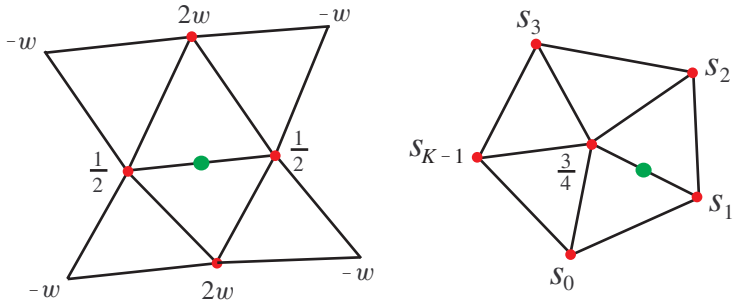



Figure 7. Weights for $e$-vertex: butterfly scheme (left), modified butterfly scheme (right).

## 3. Linear extensions

In this section, we review several extensions of stationary linear schemes for the refinement of points to stationary linear schemes which refine other objects.

**3.1. Matrix subdivision schemes.** Matrix schemes are defined by matrix masks and refine sequences of vectors. Although, in the geometric setting, the schemes of §2 refine sequences of control points in $\mathbb{R}^2$ or in $\mathbb{R}^3$, the schemes operate on each component of the vectors in the same way, such that the refinement of one component is independent of the other components. This property is very important in geometric applications, since the subdivision schemes commute with affine transformations (the schemes are *affine invariant*). The schemes presented here are not affine invariant, and their main application is in multiwavelets constructions [12], [58] and in the

analysis of multivariate subdivision schemes for control points as indicated in §2.3 (see, e.g., [24]).

A finite set of matrices of order $d \times d$, $\boldsymbol{A} = \{A_\alpha : \alpha \in \sigma(\boldsymbol{A}) \subset \mathbb{Z}^s\}$, defines a matrix subdivision scheme $S_{\boldsymbol{A}}$ with a refinement rule

$$(S_{\boldsymbol{A}} \boldsymbol{v})_\alpha = \sum_{\beta \in \mathbb{Z}^s} A_{\alpha - 2\beta} v_\beta, \quad \boldsymbol{v} = \{v_\alpha \in \mathbb{R}^d : \alpha \in \mathbb{Z}^s\}. \tag{1}$$

Given initial "control vectors" $\boldsymbol{v}^0 = \{v_\alpha^0 \in \mathbb{R}^d : \alpha \in \mathbb{Z}^s\}$, the matrix subdivision scheme $S_{\boldsymbol{A}}$ generates a sequence of control vectors by

$$\boldsymbol{v}^{k+1} = S_{\boldsymbol{A}} \boldsymbol{v}^k, \quad k = 1, 2, \ldots. \tag{2}$$

The notion of uniform convergence from §2 can be extended to this case, by considering the convergence of each of the $d$ components of the vectors. The convergence analysis has a linear algebra component to it, in addition to the analysis component. By considering the matrices

$$B_\gamma = \sum_{\beta \in \mathbb{Z}^s} A_{\gamma - 2\beta}, \quad \gamma \in E_s, \tag{3}$$

one can easily conclude a necessary condition for convergence. This condition is the analogue of condition (2.16), stating that for any initial sequence of control vectors $\boldsymbol{v}^0$, and any $x \in \mathbb{R}^s$,

$$(S_{\boldsymbol{A}}^\infty \boldsymbol{v}^0)(x) \in \operatorname{span}\{u \in \mathbb{R}^d : B_\gamma u = u \text{ for all } \gamma \in E_s\}. \tag{4}$$

In the extreme case of schemes with $B_\gamma = I$, $\gamma \in E_s$, the space in (4) is $\mathbb{R}^d$, and no condition of linear-algebra type is imposed. Such are the schemes used in the analysis of convergence and smoothness of multivariate schemes for points. Schemes for which the space in (4) is $\mathbb{R}^d$, are very similar to schemes with a scalar mask [17]. In the other extreme case, the space in (4) is one dimensional with vectors of equal components, implying that the limit vector function $S_{\boldsymbol{A}}^\infty \boldsymbol{v}^0$, has equal components. An example of this type of schemes is provided by matrix subdivision schemes generating multiple knot $B$-spline curves (see, e.g., [53]). This latter extreme case is the most relevant to the construction of multiwavelets.

In [13] and in [38], univariate ($s = 1$) matrix schemes with the space (4) of dimension $m$, $1 \leq m \leq d$, are studied. An appropriate change of basis, depending on the structure of the space (4), facilitates the extension of the factorization of scalar symbols to a certain factorization of matrix symbols. This factorization is sufficient for convergence and smoothness analysis of matrix schemes, and is also necessary under an extension of the notion of $L_\infty$-stability (see §2.3) to the matrix case. Multivariate matrix schemes with the space (4) of general dimension are considered in [56].

In the next section we discuss a special type of matrix subdivision schemes, which is relevant to curve design from locations and normals, and to the generation of

functions from the point values of the functions and their derivatives. The use of analogous schemes for the generation of surfaces from locations and normals is not straightforward, and leads to non-linear schemes.

**3.2. Hermite subdivision schemes.** The first Hermite schemes to be studied were univariate and interpolatory [51]. Interpolatory Hermite subdivision schemes are matrix schemes, such that the components of the vectors are regarded as the value of a function and its consecutive derivatives up to a certain order at the points of $2^{-k}\mathbb{Z}^s$. Non-interpolatory Hermite subdivision schemes were introduced later [43].

**3.2.1. Univariate interpolatory Hermite schemes.** The most common construction of interpolatory Hermite subdivision schemes is similar to the second construction method presented in §2.1. The approximation operator $A$ is an extension of the one in (2.9). For interpolatory schemes, it is a polynomial interpolation operator of the form

$$(Af)(x) = \sum_{\alpha\in\mathcal{A}} \sum_{i=0}^{d-1} w_{\alpha,i}(x) f^{(i)}(\alpha), \tag{5}$$

satisfying $D^i(Af)(\alpha) = f^{(i)}(\alpha)$, $\alpha \in \mathcal{A}$, $i = 0, 1, \ldots, d-1$.

The refinement is similar to (2.10), namely

$$v_{2\alpha}^{k+1} = v_\alpha^k, \qquad \left(v_{2\alpha+1}^{k+1}\right)_j = \sum_{\beta\in\mathcal{A}} \sum_{i=0}^{d-1} D^j w_{\beta,i}(1/2)(v_{\alpha+\beta}^k)_i, \quad 0 \le j \le d-1. \tag{6}$$

In (6), $(v)_i$ denotes the $i$-th component of the vector $v$. The refinement (6) can be written in terms of a matrix mask as,

$$v_\alpha^{k+1} = \sum_{\beta\in\mathbb{Z}^s} A_{\alpha-2\beta}^{(k)} v_\beta^k, \quad \alpha \in \mathbb{Z}^s, \tag{7}$$

where the matrices with even indices are

$$A_{2\alpha}^{(k)} = \delta_{\alpha,0} I_{d\times d}, \quad \alpha \in \mathbb{Z}, \tag{8}$$

with $\delta_{\alpha,0} = 0$ for $\alpha \neq 0$, and $\delta_{0,0} = 1$. The matrices with odd indices depend on the refinement level $k$, and have the form

$$A_{2\alpha+1}^{(k)} = \Lambda_d(2^k) A_{2\alpha+1}^{(0)} \Lambda_d(2^{-k}), \quad \alpha \in \mathbb{Z}, \tag{9}$$

with $\Lambda_d(h) = \operatorname{diag}(1, h, h^2, \ldots, h^{d-1})$ and

$$A_{1-2\alpha}^{(0)} = \left\{ D^i w_{\alpha,j}\left(\frac{1}{2}\right) \right\}_{i,j,=0}^{d-1}, \quad \alpha \in \mathcal{A}. \tag{10}$$

The powers of 2 in (9) are due to the fact that derivatives of polynomials are not scale invariant. More precisely if $q(x) = p(hx)$, with $p$ a polynomial, then $(D^j q)(hx_0) = h^j (D^j p)(x_0)$.

An interpolatory Hermite scheme is termed *uniformly convergent* if there is a limit vector function $F$ of the form $F = (D^j f,\ 0 \le j \le d-1)^T$, with $f \in C^{d-1}(\mathbb{R})$, satisfying for any closed interval $[a, b]$,

$$\lim_{k\to\infty} \sup_{\alpha\in 2^k[a,b]\cap\mathbb{Z}} \left\| F(2^{-k}\alpha) - v_\alpha^k \right\| = 0,$$

with $\|\cdot\|$ any norm in $\mathbb{R}^d$.

**Example (a two-point Hermite interpolatory scheme).** The scheme is given by the non-zero matrices of its mask:

$$A_0 = I_{2\times 2}, \quad A_1^{(k)} = \begin{pmatrix} \frac{1}{2} & \nu 2^{-k} \\ -\mu 2^k & \frac{1-\mu}{2} \end{pmatrix}, \quad A_{-1}^{(k)} = \begin{pmatrix} \frac{1}{2} & -\nu 2^{-k} \\ \mu 2^k & \frac{1-\mu}{2} \end{pmatrix}.$$

This scheme with $\nu = 1/8$ and $\mu = 3/2$ generates the $C^1$ piecewise Hermite cubic interpolant to the data $\{v_i^0 = (f(i), f'(i))^T : i \in \mathbb{Z}\}$, while for $\nu = 0$, $\mu = 1$, it generates the piecewise linear interpolant to the given function's values at the integers, which is only $C^0$. By the analysis to be reviewed, it can be shown that for $0 < \nu < 1/4$, $\mu = 4\nu + 1$, the limit functions generated by the scheme are $C^1$. (See, e.g. [36].)

One method for the convergence analysis of such schemes is based on deriving an equivalent stationary matrix scheme, refining vectors of $(d-1)$-th order divided differences, obtained from the original control vectors. The limit of such a scheme, if it exists, necessarily consists of equal components, which are the derivative of order $d-1$ of the smooth function $f$ [37].

More precisely, the divided difference vector $u_n^k$ at level $k$ is defined for each $n \in \mathbb{Z}$ by

$$(u_n^k)_j = [\tau_{j+1}, \tau_{j+2}, \ldots, \tau_{j+d}]f, \quad j = 0, \ldots, d-1,$$

with $\tau_1 = \cdots = \tau_{d-1} = (n-1)2^{-k}$, $\tau_d = \tau_{d+1} = \cdots = \tau_{2d-1} = n2^{-k}$. Here we use the definition of divided differences, allowing repeated points for functions with enough derivatives (see, e.g., [6, Chapter 1]). In our setting all integer points have multiplicity $d$. The vector $u_n^k$ can be derived from the vectors $v_{n-1}^k$ and $v_n^k$.

The symbol $D(z)$ of the matrix scheme refining the control vectors $\boldsymbol{u}^k = \{u_n^k : n \in \mathbb{Z}\}$ can be obtained recursively from the symbol $D^{[0]}(z) = \sum_\alpha A_\alpha^{(0)} z^\alpha$, by algebraic manipulations, involving multiplication by certain matrix Laurent polynomials and their inverses.

It is proved in [37] that the matrix symbol $D(z)$ is a matrix Laurent polynomial if the scheme (7) reproduces polynomials of degree $\le d-2$, and that necessarily a scheme of the form (7), which generates $C^{d-1}$ functions, reproduces polynomials of degree $\le d-1$. In (5), the degree of the interpolation polynomial is $d|\mathcal{A}| - 1$,

so the scheme (7), with the mask (8),(9),(10), reproduces polynomials of degree at least $2d-1$, as $\mathcal{A}$ contains at least the points $0, 1$. These arguments lead to the conclusion that the Hermite subdivision scheme $S_{\boldsymbol{A}}$, refining the control vectors $\boldsymbol{v}^k$ can be transformed into the matrix subdivision scheme $S_{\boldsymbol{D}}$ for the control vectors $\boldsymbol{u}^k$.

To determine the convergence of the scheme $S_{\boldsymbol{D}}$, which is equivalent to the convergence of the original Hermite subdivision scheme $S_{\boldsymbol{A}}$ to $C^{d-1}$ functions, we observe that the component $(u_n^k)_j$, in case of convergence, approximates $f^{(d-1)}(2^{-k}n)$ for $j = 1, \ldots, d$. Thus as in the case of control points, a necessary condition for convergence is the contractivity of the scheme which refines the differences $(u_n^k)_j - (u_n^k)_{j-1}$, $j = 2, \ldots, d$, $(u_n^k)_1 - (u_{n-1}^k)_d$, $n \in \mathbb{Z}$. Indeed, such a scheme exists, and its symbol is a matrix Laurent polynomial when (7) reproduces polynomials of degree $\leq d-1$ [37], guaranteeing that the contractivity of this scheme can be checked by algebraic manipulations.

The analysis of higher order smoothness is along the same lines.

**3.3. *B*-spline subdivision schemes for compact sets.** In the last years, the univariate $B$-spline schemes were extended to operate on data consisting of compact sets [27], [28]. The motivation for the study of such schemes is the problem of approximating a 3D object from a discrete set of its 2D parallel cross-sections, and the problem of approximating a 2D shape from a discrete set of its 1D parallel cross-sections. In both problems, either the 3D object or the 2D shape is regarded as a univariate set-valued function, with its parallel cross-sections as images. The $B$-spline subdivision schemes are adapted to this setting, so that the limit set-valued function generated by the subdivision from samples of a continuous set-valued function, approximates it.

For initial data $\boldsymbol{F}^0 = \{F_i^0 \subset \mathbb{R}^n : i \in \mathbb{Z}\}$ consisting of convex compact sets, averages of numbers in the execution of a scheme, can be replaced by Minkowski averages of sets. A *Minkowski average* of sets $B_1, \ldots, B_\ell \subset \mathbb{R}^n$ with weights $\lambda_1, \ldots, \lambda_\ell \in \mathbb{R}$, $\sum_{i=1}^{\ell} \lambda_i = 1$, is the set

$$\sum_{i=1}^{\ell}{}_M \lambda_i B_i = \Big\{ \sum_{i=1}^{\ell} \lambda_i b_i : b_i \in B_i \Big\}.$$

Thus the $m$-th degree $B$-spline subdivision scheme (1.4) can be adapted to convex compact sets by the refinement rule

$$F_i^{k+1} = \sum_j{}_M a_{i-2j}^{[m]} F_j^k, \quad i \in \mathbb{Z}, \tag{11}$$

with $\boldsymbol{a}^{[m]} = \{a_i^{[m]}, i = 0, \ldots, m+1\}$ given in (1.2). Since the coefficients of the mask are positive, the sets $\boldsymbol{F}^k$, $k \geq 1$, generated by the subdivision scheme $S_{M,a^{[m]}}$ with the refinement rule (11) are compact and convex [27]. By the associativity and distributivity of the Minkowski average of convex sets with positive weights, it can

be deduced straightforwardly that the limit generated by $S_{M,a^{[m]}}$ from $\boldsymbol{F}^0$, when $\boldsymbol{F}^0$ consists of convex compact sets, is

$$(S^\infty_{M,\boldsymbol{a}^{[m]}}\boldsymbol{F}^0)(t) = \sum_{i\in\mathbb{Z}_M} F^0_i B_m(t-i). \tag{12}$$

In (12) the convergence is in the Hausdorff metric, defined for two sets $A$, $B$ in $\mathbb{R}^n$, by

$$\text{haus}(A,B) = \max\Big\{\sup_{a\in A}\inf_{b\in B}\|a-b\|,\ \sup_{b\in B}\inf_{a\in A}\|a-b\|\Big\}$$

with $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^n$.

The subdivision scheme $S_{M,a^{[m]}}$ has approximation properties. It is shown in [27] that for a set-valued function $G$ with convex compact images, which is Lipschitz continuous, namely satisfies $\text{haus}(G(t), G(t+\Delta)) = O(\Delta t)$, and for initial data $\boldsymbol{F}^0_h = \{G(ih) : i\in\mathbb{Z}\}$

$$\text{haus}\left((S^\infty_{M,\boldsymbol{a}^{[m]}}\boldsymbol{F}^0_h)(t), G(t)\right) = O(h). \tag{13}$$

The subdivision $S_{M,a^{[m]}}$ fails to approximation set-valued functions with general compact images. As is shown in [29], for initial data $\boldsymbol{F}^0$ consisting of general compact sets,

$$S^\infty_{M,\boldsymbol{a}^{[m]}}\boldsymbol{F}^0 = \sum_{i\in\mathbb{Z}}\langle F^0_i\rangle B_m(\cdot - i)$$

with $\langle F^0_i\rangle$ the convex hull of $F^0_i$. Thus $S^\infty_{M,\boldsymbol{a}^{[m]}}\boldsymbol{F}^0$ is convex even when the initial sets are non-convex, and it cannot approximate set-valued functions with general compact sets as images.

There is another adaptation of the $B$-spline subdivision schemes to compact sets [28], which yields approximation also in case of set-valued functions with general compact sets as images. This adaptation is obtained by using the first construction in §2.1 for $s=1$, and by replacing the average of two numbers by the *metric average* of two compact sets, introduced in [1],

$$A\oplus_t B = \left\{ta + (1-t)b : (a,b)\in\Pi(A,B)\right\}$$

with

$$\Pi(A_0,A_1) = \Big\{(a_0,a_1) : a_i\in A_i,\ i=0,1,\\ \|a_0-a_1\| = \min_{a\in A_j}\|a_i-a\|,\ j=1-i,\ \text{for } i=0 \text{ or } 1\Big\}.$$

The refinement rule of the resulting scheme $S_{MA,m}$ is achieved by the $m+1$ steps,

$$\begin{aligned} F^{k+1,0}_{2i} &= F^k_i, \quad F^{k+1,0}_{2i+1} = F^k_i, \quad i\in\mathbb{Z},\\ F^{k+1,j}_i &= F^{k+1,j-1}_i \oplus_{\frac{1}{2}} F^{k+1,j-1}_{i-1}, \quad i\in\mathbb{Z},\ j=1,\dots,m\\ F^{k+1}_i &= F^{k+1,m}_i, \quad i\in\mathbb{Z} \end{aligned} \tag{14}$$

The refinement rule (14) is denoted formally by $\boldsymbol{F}^{k+1} = S_{MA,m}\boldsymbol{F}^k$.

Two important properties of the metric average, which are central to its application in $B$-spline subdivision schemes are

$$A \oplus_t A = A, \quad \text{haus}(A \oplus_t B, A \oplus_s B) = |s-t|\,\text{haus}(A, B), \tag{15}$$

for $(s, t) \in [0, 1]$.

Let the sequence $\{H^k\}$ consist of the "piecewise linear" set valued functions, interpolating $\{\boldsymbol{F}^k = S^k_{MA,m}\boldsymbol{F}^0\}$,

$$H^k(t) = F^k_i \oplus_{\lambda(t)} F^k_{i+1}, \quad 2^{-k}i \le t < 2^{-k}(i+1),\ i \in \mathbb{Z},\ k = 0, 1, 2, \ldots, \tag{16}$$

with $\lambda(t) = i + 1 - 2^k t$. It is proved in [28], with the aid of the metric property of the metric average (the second equality in (15)) and the completeness of the metric space of compact sets with the Hausdorff metric, that the sequence $\{H^k(t)\}$ converges to a limit set-valued function denoted by $S^\infty_{MA,m}\boldsymbol{F}^0$.

Moreover, for $G$ a Lipschitz continuous set valued function with general compact sets as images, the limit generated by the scheme $S_{MA,m}$ starting from $\boldsymbol{F}^0_h = \{G(ih) : i \in \mathbb{Z}\}$ approximates $G$ with "error" given by

$$\text{haus}\left((S^\infty_{MA,m}\boldsymbol{F}^0_h)(t), G(t)\right) = O(h), \quad t \in \mathbb{R}. \tag{17}$$

**3.4. A blending-based subdivision scheme for nets of curves.** The quadratic $B$-spline scheme (Chaikin algorithm) was extended to the refinement of nets of curves in [15]. A net of curves with parameter $d > 0$ consists of two families of continuous curves

$$\left\{\phi_i(s) : 0 \le i \le n,\ s \in [0, md]\right\}, \quad \left\{\psi_j(t) : 0 \le j \le m,\ t \in [0, nd]\right\}$$

satisfying the compatibility condition

$$\phi_i(jd) = \psi_j(id), \quad i = 0, \ldots, n,\ j = 0, \ldots, m.$$

Such a net is denoted by $\mathcal{N}(d, \{\phi_i\}_{i=0}^n, \{\psi_j\}_{j=0}^m)$. The *blending-based Chaikin-type* scheme refines a net of curves, $\mathcal{N}_0 = \mathcal{N}(d, \{\phi^0_i\}_{i=0}^n, \{\psi^0_j\}_{j=0}^m)$ into a net of curves $\mathcal{N}_1 = \mathcal{N}\left(\frac{d}{2}, \{\phi^1_i\}_{i=0}^{2n-1}, \{\psi^1_j\}_{j=0}^{2m-1}\right)$. A repeated application of such refinements generates a sequence of nets $\left\{\mathcal{N}_k = \mathcal{N}\left(\frac{d}{2^k}, \{\phi^k_i\}_{i=0}^{n_k}, \{\psi^k_j\}_{j=0}^{m_k}\right) : k \in \mathbb{Z}_+\right\}$, with $n_k = 2^k(n-1)+1, m_k = 2^k(m-1)+1$, which converges uniformly to a continuous surface [15].

The construction of the refinement rule is analogous to the second method in §2.1. The approximation operator $A$ maps a net of curves $\mathcal{N}(d, \{\phi_i\}_{i=0}^n, \{\psi_j\}_{j=0}^m)$ into the piecewise Coons patch surface, interpolating the curves of the net,

$$\begin{gathered}\mathcal{C}(\mathcal{N})(s,t) = C(\phi_i, \phi_{i+1}, \psi_j, \psi_{j+1}; d)(s - jd, t - id),\\ (s,t) \in [jd, jd+d] \times [id, id+d], \quad i = 0, \ldots, n-1,\ j = 0, \ldots, m-1,\end{gathered} \tag{18}$$

with $C(\phi_i, \phi_{i+1}, \psi_j, \psi_{j+1}; d)$ a Coons patch [16].

Four continuous curves $\phi_0, \phi_1, \psi_0, \psi_1$ defined on $[0, h]$ and satisfying $\phi_i(jh) = \psi_j(ih)$, $i, j = 0, 1$, define a Coons patch on $[0, h]^2$. For $(s, t) \in [0, h]^2$ the Coons patch is given by

$$\begin{aligned} &C(\phi_0, \phi_1, \psi_0, \psi_1; h)(s, t) \\ &\quad = \left[\left(1 - \frac{t}{h}\right)\phi_0(s) + \frac{t}{h}\phi_1(s)\right] + \left[\left(1 - \frac{s}{h}\right)\psi_0(t) + \frac{s}{h}\psi_1(t)\right] \\ &\qquad - \left[\left(1-\frac{s}{h}\right)\left(\left(1-\frac{t}{h}\right)\phi_0(0)+\frac{t}{h}\phi_1(0)\right)+\frac{s}{h}\left(\left(1-\frac{t}{h}\right)\phi_0(h)+\frac{t}{h}\phi_1(h)\right)\right]. \end{aligned} \tag{19}$$

The Coons patch is blending between two surfaces. One is interpolating linearly between corresponding points of $\phi_0, \phi_1$ and the other between the corresponding points of $\psi_0, \psi_1$. (These two surfaces are the two first terms on the right-hand side of (19)). It is easy to verify that $C(\phi_0, \phi_1, \psi_0, \psi_1; h)$ coincides with the four curves on the boundary of $[0, h]^2$, namely that

$$\begin{aligned} C(\phi_0, \phi_1, \psi_0, \psi_1; h)(jh, t) &= \psi_j(t), \quad j = 0, 1, \\ C(\phi_0, \phi_1, \psi_0, \psi_1; h)(s, ih) &= \phi_i(s), \quad i = 0, 1. \end{aligned}$$

Regarding the Coons patch of four curves as the analogue of a linear segment between two points, the Chaikin scheme for the refinement of control points is "extended" to nets of curves, by sampling each of the Coons patches of $\mathcal{C}(\mathcal{N}_k)$ at $1/4$ and $3/4$ of the corresponding parameters values. Thus the refinement rule analogous to (2.11) is

$$\phi_{2i}^{k+1}(s) = \mathcal{C}(\mathcal{N}_k)\left(s, \left(i + \frac{1}{4}\right)\frac{d}{2^k}\right), \quad \phi_{2i+1}^{k+1} = \mathcal{C}(\mathcal{N}_k)\left(s, \left(i + \frac{3}{4}\right)\frac{d}{2^k}\right), \quad i = 0, \ldots, n_k - 1, \tag{20}$$

$$\psi_{2j}^{k+1}(t) = \mathcal{C}(\mathcal{N}_k)\left(\left(j + \frac{1}{4}\right)\frac{d}{2^k}, t\right), \quad \psi_{2j+1}^{k+1} = \mathcal{C}(\mathcal{N}_k)\left(\left(j + \frac{3}{4}\right)\frac{d}{2^k}, t\right), \quad j = 0, \ldots, m_k - 1. \tag{21}$$

This refinement rule generates a refined net of curves after a simple reparametrization. This is written formally as $\mathcal{N}_{k+1} = S_{BC}\mathcal{N}_k$.

The proof of convergence of the scheme $S_{BC}$ is not an extension of the analysis of §2.3, but is based on the proximity of $S_{BC}$ to a new subdivision scheme $S_{\boldsymbol{a}}$ for points, which is proved to be convergent by the analysis of §2.3.

Convergence proofs by proximity to linear stationary schemes for points are employed, e.g., in [35] for the analysis of linear non-stationary schemes, and in [60] for the analysis of a certain class of non-linear schemes.

Another important ingredient in the convergence proof is a property of a net of curves, which is preserved during the refinements with $S_{BC}$. A net of curves

$\mathcal{N}(d, \{\phi_i\}_{i=0}^n, \{\psi_j\}_{j=0}^m)$ is said to have the *M-property* if the second divided differences of all curves of the net at three points restricted to intervals of the form $\left[\ell d, \left(\ell + \frac{1}{2}\right)d\right]$, $\ell \in (1/2)\mathbb{Z}$ in the domain of definition of the curves, are all bounded by a constant $M$.

The sequence $\{\mathcal{C}(\mathcal{N}_k) : k \in \mathbb{Z}_+\}$ of continuous surfaces is shown to be a Cauchy sequence for $\mathcal{N}_0$ with the $M$-property, by comparison of one refinement of $S_{BC}$ with one refinement of $S_{\boldsymbol{a}}$. The scheme $S_{\boldsymbol{a}}$ is constructed to be in proximity to $S_{BC}$ in the sense that

$$\left\| \mathcal{E}(S_{BC}\mathcal{N}_k) - S_{\boldsymbol{a}}(\mathcal{E}(\mathcal{N}_k)) \right\| \leq \frac{3}{2} M \left( \frac{d}{2^{k+1}} \right)^2, \tag{22}$$

with $\mathcal{E}(\mathcal{N}_k) = \left\{ C(\mathcal{N}_k)\left(i\frac{d}{2}, j\frac{d}{2}\right),\ 0 \leq i \leq 2m_k,\ 0 \leq j \leq 2n_k \right\}$, and with $M$ the constant in the $M$-property satisfied by all the nets $\{\mathcal{N}_k : k \in \mathbb{Z}_+\}$ which are generated by $S_{BC}$.

Although the limit of the Cauchy sequence $\{\mathcal{C}(\mathcal{N}_k) : k \in \mathbb{Z}_+\}$ is only $C^0$, it is conjectured in [15] that $S_{BC}$ generates $C^1$ surfaces from initial curves which are $C^1$. This conjecture is based on simulations.

**Acknowledgement.** The author wishes to thank David Levin and Adi Levin for helping with the figures and the references.

## References

[1] Artstein, Z., Piecewise linear approximations of set-valued maps. *J. Approx. Theory* **56** (1989), 41–47.

[2] Ball, A. A., Storry, D. J. T., Conditions for tangent plane continuity over recursively generated b-spline surfaces. *ACM Transactions on Graphics* **7** (1988), 83–102.

[3] Ball, A. A., Storry, D. J. T., Design of an $n$-sided surface patch from Hermite boundary data. *Comput. Aided Geom. Design* **6** (1989), 111–120.

[4] de Boor, C., Cutting corners always works. *Comput. Aided Geom. Design* **4** (1987), 125–131.

[5] de Boor, C., Quasi-interpolants and approximation power of multivariate splines. In *Computation of Curves and Surfaces* (ed. by W. Dahmen, M. Gasca, C. Micchelli), NATO ASI Series, Kluwer Academic Press, 1990, 313–346.

[6] de Boor, C., *A Practical Guide to Splines*. Appl. Math. Sci. 27, Springer-Verlag, New York 2001.

[7] de Boor, C., Höllig, K., Riemenschneider, S., *Box Splines*. Appl. Math. Sci. 98, Springer-Verlag, New York 1993.

[8] Catmull, E., Clark, J., Recursively generated b-spline surfaces on arbitrary topological meshes. *Comput. Aided Design* **10** (1978), 350–355.

[9] Cavaretta, A. S., Dahmen, W., Micchelli, C. A., *Stationary Subdivision*. Mem. Amer. Math. Soc. 452, Amer. Math. Soc., Providence, RI, 1991.

[10] Chaikin, G. M., An algorithm for high speed curve generation. *Computer Graphics and Image Processing* **3** (1974), 346–349.

[11] Cirak, R., Scott, M., Schröder, P., Ortiz, M., Antonsson, E., Integrated modeling, finite-element analysis and design for thin-shell structures using subdivision. *Comput. Aided Design* **34** (2002), 137–148.

[12] Cohen, A., Daubechies, I., Plonka, G., Regularity of refinable function vectors. *J. Fourier Anal. Appl.* **3** (1997), 295–324.

[13] Cohen, A., Dyn, N., Levin, D., Stability and inter-dependence of matrix subdivision schemes. In *Advanced Topics in Multivariate Approximation* (ed. by F. Fontanella, K. Jetter, P. J. Laurent), Ser. Approx. Decompos. 8, World Scientific Publishing Co., River Edge, NJ, 1996, 33–45.

[14] Cohen, E., Lyche, T., Riesenfeld, R. F., Discrete b-splines and subdivision techniques in Computer-Aided Geometric Design and Computer graphics. *Computer Graphics and Image Processing* **14** (1980), 87–111.

[15] Conti, C., Dyn, N., Blending-based Chaikin-type subdivision schemes for nets of curves. In *Mathematical Methods for Curves and Surfaces: Tromso 2004* (ed. by M. Dahlen, K. Morken, L. Schumaker), Mod. Methods Math., Nashboro Press, Brentwood, TN, 2005, 51–68.

[16] Coons, S. A., Surface for computer aided design. Tech. Rep., MIT, 1964.

[17] Cotronei, M., Sauer, T., Full rank filters and polynomial reproduction. Preprint.

[18] Daubechies, I., *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[19] Daubechies, I., Lagarias, J. C., Two-scale difference equations I, existence and global regularity of solutions. *SIAM J. Math. Anal.* **22** (1992), 1388–1410.

[20] Daubechies, I., Lagarias, J. C., Two-scale difference equations II, local regularity, infinite products of matrices and fractals. *SIAM J. Math. Anal.* **23** (1992), 1031–1079.

[21] DeRose, T., Kass, M., Truong, T., Subdivision surfaces in character animation. In *SIGGRAPH 98 Conference Proceedings*, Annual Conference Series, ACM SIGGRAPH, 1998, 85–94.

[22] Deslauriers, G., Dubuc, S., Symmetric iterative interpolation processes. *Constr. Approx.* **5** (1989), 49–68; Erratum ibid. **8** (1992), 125–126.

[23] Doo, D., Sabin, M., Analysis of the behaviour of recursive division surface near extraordinary points. *Comput. Aided Design* **10** (1978), 356–360.

[24] Dyn, N., Subdivision schemes in computer aided geometric design. In *Advances in Numerical Analysis II, Wavelets Subdivision Algorithms and Radial Basis Functions* (ed. by W. A. Light), Oxford Sci. Publ., Oxford University Press, New York 1992, 36–104.

[25] Dyn, N., Subdivision: Analysis of convergence and smoothness by the formalism of Laurent polynomials. In *Tutorials on Multiresolution in Geometric Modelling* (ed. by M. Floater, A. Iske, E. Quak), Math. Vis., Springer-Verlag, Berlin 2002, 51–68.

[26] Dyn, N., Three families of nonlinear subdivision schemes. In *Multivariate Approximation and Interpolation* (ed. by K. Jetter, M. D. Buhmann, W. Haussmann, R. Schaback, J. Stöckler), Elsevier, 2005,23–38.

[27] Dyn, N., Farkhi, E., Spline subdivision schemes for convex compact sets. *J. Comput. Appl. Math.* **119** (2000), 133–144.

[28] Dyn, N., Farkhi, E., Spline subdivision schemes for compact sets with metric averages. In *Trends in Approximation Theory* (ed. by K. Kopotun, T. Lyche, M. Neamtu), Innov. Appl. Math., Vanderbilt University Press, Nashville, TN, 2001, 93–102.

[29] Dyn, N., Farkhi, E., Set-valued approximations with Minkowski averages – convergence and convexification rates. *Numer. Funct. Anal. Optim.* **25** (2004), 363–377.

[30] Dyn, N., Floater, M., Hormann, K., A $C^2$ four-point subdivision scheme with fourth order accuracy and its extensions. In *Mathematical Methods for Curves and Surfaces: Tromso 2004* (ed. by M. Dahlen, K. Morken, L. Schumaker), Mod. Methods Math., Nashboro Press, Brentwood, TN, 2005, 145–156.

[31] Dyn, N., Gregory, J. A., Levin, D., A four-point interpolatory subdivision scheme for curve design. *Comput. Aided Geom. Design* **4** (1987), 257–268.

[32] Dyn, N., Gregory, J. A., Levin, D., A butterfly subdivision scheme for surface interpolation with tension control. *ACM Transactions on Graphics* **9** (1990), 160–169.

[33] Dyn, N., Gregory, J. A., Levin, D., Analysis of uniform binary subdivision schemes for curve design. *Constr. Approx.* **7** (1991), 127–147.

[34] Dyn, N., Levin, D., Interpolating subdivision schemes for the generation of curves and surfaces. In *Multivariate Approximation and Interpolation* (ed. by W. Haussmann, K. Jetter), Internat. Ser. Numer. Math. 94, Birkhäuser Verlag, Basel 1990, 91–106.

[35] Dyn, N., Levin, D., Analysis of asymptotically equivalent binary subdivision schemes. *J. Math Anal Appl* **193** (1995), 594–621.

[36] Dyn, N., Levin, D., Analysis of Hermite-type subdivision schemes. In *Approximation Theory VIII – Wavelets and Multilevel Approximation* (ed. by C. Chui, L. Schumaker), Vol. 2, Ser. Approx. Decompos. 6, World Scientific Publications, River Edge, NJ, 1995, 117–124.

[37] Dyn, N., Levin, D., Analysis of Hermite-interpolatory subdivision schemes. In *Spline Functions and the Theory of Wavelets* (ed. by S. Dubuc), Centre de Recherches Mathématiques, CRM Proc. Lecture Notes 18, Amer. Math. Soc., Providence, RI, 1999, 105–113.

[38] Dyn, N., Levin, D., Matrix subdivision-analysis by factorization. In *Approximation Theory: A volume dedicated to Blagovest Sendov* (ed. by B. Bojanov), Darba, Sofia 2002, 187–211.

[39] Dyn, N., Levin, D., Subdivision schemes in geometric modelling. *Acta Numer.* **11** (2002), 73–144.

[40] Gregory, J. A., Qu, R., Non-uniform corner cutting. *Comp. Aided Geom. Design* **13** (8) (1996), 763–772.

[41] Halstead, M., Kass, M., DeRose, T., Efficient, fair interpolation using catmull-clark sutfaces. In *SIGGRAPH 93 Conference Proceedings*, Annual Conference Series, ACM SIGGRAPH, 1993, 35–44.

[42] Han, B., Computing the smoothness exponent of a symmetric multivariate refinable function. *SIAM J. Matrix Anal. Appl.* **24** (2003), 693–714.

[43] Han, B., Yu, T. P., Xue, Y., Noninterpolatory Hermit subdivision schemes. *J. Math. Comp.* **74** (2005), 1345–1367.

[44] Hed, S., Analysis of subdivision schemes for surfaces. Master Thesis, Tel-Aviv University, 1990.

[45] Jia, R. Q., Characterization of smoothness of multivariate refinable functions in Sobolev spaces. *Trans. Amer. Math. Soc.* **351** (1999), 4089–4112.

[46] Karĉiauskas, K., Peters, J., Reif, U., Shape characterization of subdivision surfaces: case studies. *Comput. Aided Geom. Design* **21** (2004), 601–614.

[47] Kobbelt, L., Interpolatory subdivision on open quadrilateral nets with arbitrary topology. *Computer Graphics Forum* **15** (1996), 409–420.

[48] Kobbelt, L., Sqrt(3) subdivision. In *Proceedings of SIGGRAPH 2000*, Annual Conference Series, ACM-SIGGRAPH, 2000, 103–112.

[49] Kobbelt, L., Hesse, T., Prautzsch, H., Schweizerhof, K., Interpolatory subdivision on open quadrilateral nets with arbitrary topology. *Computer Graphics Forum* **15** (1996), 409–420.

[50] Loop, C., Smooth spline surfaces based on triangles. Master Thesis, University of Utah, Department of Mathematics, 1987.

[51] Merrien, J. L., A family of Hermite interpolants by bisection algorithms. *Numer. Algorithms* **2** (1992), 187–200.

[52] Peters, J., Reif, U., Shape characterization of subdivision surfaces-basic principles. *Comput. Aided Geom. Design* **21** (2004), 585–599.

[53] Plonka, G., Approximation order provided by refinable function vectors. *Constr. Approx.* **13** (1997), 221–244.

[54] de Rahm, G., Sur une courbe plane. *J. Math. Pures Appl.* (9) **35** (1956), 25–42.

[55] Reif, U., A unified approach to subdivision algorithms near extraordinary points. *Comput. Aided Geom. Design* **12** (1995), 153–174.

[56] Sauer, T., Stationary vector subdivision – quotient ideals, differences and approximation power. *RACSAM Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat.* **96** (2002), 257–277.

[57] Shenkman, P., Computing normals and offsets of curves and surfaces generated by subdivision schemes. Master Thesis, Tel-Aviv university, 1996.

[58] Strella, V., Multiwavelets: Theory and Applications. PhD Thesis, MIT, 1996.

[59] Velho, L., Zorin, D., 4-8 subdivision. *Comput. Aided Geom. Design* **18** (2001), 397–427.

[60] Wallner, J., Dyn, N., Convergence and $C^1$ analysis of subdivision schemes on manifolds by proximity. *Comput. Aided Geom. Design* **22** (2005), 593–622.

[61] Zorin, D., Schröder, P., Sweldens, W., Interpolating subdivision for meshes with arbitrary topology. In *SIGGRAPH 96 Conference Proceedings*, Annual Conference Series, ACM-SIGGRAPH, 1996, 189–192.

School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel
E-mail: niradyn@post.tau.ac.il

# Wave propagation software, computational science, and reproducible research

Randall J. LeVeque*

*Dedicated to Germund Dahlquist (1925–2005) and Joseph Oliger (1941–2005), two of the influential mentors who have shaped my career. They each inspired generations of students with their interest in the connections between mathematics and computation.*

**Abstract.** Wave propagation algorithms are a class of high-resolution finite volume methods for solving hyperbolic partial differential equations arising in diverse applications. The development and use of the CLAWPACK software implementing these methods serves as a case study for a more general discussion of mathematical aspects of software development and the need for more reproducibility in computational research. Sample applications discussed include medical applications of shock waves and geophysical fluid dynamics modeling volcanoes and tsunamis.

**Mathematics Subject Classification (2000).** Primary 65Y15; Secondary 74S10.

**Keywords.** Hyperbolic partial differential equations, software, numerical analysis, reproducible research, scientific computing, CLAWPACK.

## 1. Introduction

I will ultimately describe a class of numerical methods for solving hyperbolic partial differential equations, software that implements these methods, and some scientific applications. However, for the broad audience that I am honored to address in these proceedings, I would like to first make some more general comments on the topic of software development and its relation to mathematics, and on computational science and reproducible research.

I begin with a quote from a 1995 paper by J. B. Buckheit and D. L. Donoho [13] about wavelet analysis and a software package they developed to aid in studying and applying their methods:

> An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

*This research was supported in part by several NSF and DOE grants over the past decade, including NSF grants DMS-9803442, DMS-0106511, and CMS-0245206 and DOE grants DE-FG03-96ER25292, DE-FG03-00ER25292, and DE-FC02-01ER25474.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

They present this as a slogan to distill the insights of Jon Claerbout, an exploration geophysicist who has been a pioneer in this direction since the early 90s (e.g., [59]), and they give many compelling examples to illustrate its truth. I first ran across this quote on the webpage [11] of the book [12], which provides complete codes in several languages for the solutions to each of the 100-digit challenge problems proposed by Trefethen [63]. (This is set of ten computational problems, each easy to state and with a single number as the answer. The challenge was to compute at least 10 digits of each number.) In spite of some progress in the direction of reproducible research, many of the complaints of Buckheit and Donoho still ring true today, as discussed further in the recent paper by Donoho and Huo [24].

Much of my work over the past 10 years has been devoted to trying to make it easier for myself, my students, and other researchers to perform computational scholarship in the field of numerical methods for hyperbolic PDEs, and also I hope in a variety of applications areas in science and engineering where these methods are used. This work has resulted in the CLAWPACK software [40]. This software is apparently being fairly widely used, both in teaching and research. More than 5 000 people have registered to download the code, mentioning all sorts of interesting problems they plan to tackle with it. However, I am not convinced that it is being used to the extent possible in advancing scholarship of the type described above. One goal of this paper is to encourage researchers (myself included) to work harder towards this end, in my field and more generally in computational science.

I will make a distinction between software and computer programs. In my notation, *software* means a package of computer tools that are designed to be applied with some generality to a class of problems that may arise in many different applications. A *computer program* is a code that solves one particular problem. A program may be written entirely from scratch or it may employ one or more software packages as tools. This is an important distinction to make since one has different goals in developing software than in writing a specific computer program. Software is intended to be used by others and to be as general as practically possible. A program is written to solve a problem and often the author does not intend for it to be seen or used by anyone else.

Pure mathematicians search for abstract structures that transcend particular examples and that often unify disparate fields. They produce theorems that apply as generally as possible and that can be used as solid and dependable building blocks for future work. In addition to developing new algorithms, some computational mathematicians also produce theorems, rigorous results guaranteeing that a particular algorithm converges or bounds on the magnitude of the error, for example. Such theorems give us the confidence to apply the algorithm to real world problems.

Other computational mathematicians focus on the development of software that implements an algorithm in a dependable manner. This is perhaps an under-appreciated art in the mathematical world, compared to proving theorems, but I believe it is an analogous mathematical activity in many ways. In both cases the goal is to distill the essence of a set of particular examples into a general result, something that applies as broadly as possible while giving an interesting and nontrivial result that

can be built upon and used as a "subroutine" in future work. In both cases the result is an encapsulation of a set of knowledge that has well defined inputs and outputs and is believed to be proved correct, and that applies in many situations.

The process of developing a novel algorithm and writing software to implement it is also in some ways similar to the process involved in proving a theorem. One needs some mathematical insight to get started, but then working through many technical details is typically required to make everything fit together in a manner that produces the desired result. This part is not very glamorous but is a crucial part of the scholarship. Often frustrations arise when one little detail does not work out quite the way one hoped. Sometimes algorithms, like partially completed proofs, must be shelved for years until someone else makes a breakthrough in a related area that suddenly makes everything come together.

And everything does have to fit together just right; having a nice idea that seems like it should work is not enough. Glossing over the details is not allowed, and is particularly hard to pull off in a computer program. While it may be possible to slip things by the referees in the description of an algorithm in a paper (as also sometimes happens in a shoddy proof), computers will not parse the command "and then a miracle occurs". We are forced to be explicit in every step.

Of course even once a program does work, in the sense of compiling without errors and producing results that seem reasonable, we are faced with the thorny issue of "proving" that it is in fact correct. In computer science there is a whole field devoted to developing methodologies for formally proving the correctness of computer programs. In computational science the programs are often so complex and the problem it is designed to solve so ill-defined that formal correctness proofs are generally impossible. Instead the buzzwords are *Verification and Validation* (V&V for short). These can be summarized by the following mnemonic:

Verification: Are we solving the problem right?
Validation: Are we solving the right problem?

For a physical experiment modeled by partial differential equations, for example, we must verify that the computer program solves the PDEs accurately. A computational scientist must also validate the code by comparing it against experiments to ensure that the PDEs discretized are actually a sufficient model of reality for the situation being modeled. The Euler equations of gas dynamics are sufficient in some situations, but completely inadequate in other cases where viscosity plays a significant role.

Researchers in numerical analysis and scientific computing (as defined in the next section) are generally most concerned with verification, while scientists, engineers, and applied mathematicians focusing on mathematical modeling must also be concerned with validation. Even the relatively simple task of verifying that a code solves the given equations properly can be a real challenge, however, and is often as much an art as a science. A good test problem that captures the essence of some potential difficulty while having a solution that can be checked is often hard to come by, and developing test suites for different classes of algorithms is valuable scholarship in

itself. Numerous papers have been written on the subject of how best to test computer programs or software for scientific computing; see [17], [29], [33], [38], [53], or [56] for just a few approaches.

Elegance is valued in algorithm and software design as it is in other mathematical endeavors. Often the first attempt at an algorithm is not very clean; it is a brute force approach that gets the job done. Later work is often devoted to cleaning things up, perhaps in fundamental ways that greatly reduce the computational complexity, but also often in more subtle ways that simply make it more "elegant", a hard to define property that we recognize when we see.

Perhaps I am straying too far from the topic of reproducible scientific computing, but to make progress in this direction I think it is important to recognize software development as a valid and challenging mathematical activity. It takes a slightly different type of mathematician to develop the necessary intuition and skills to excel at this than is required to prove theorems, just as doing algebra vs. analysis takes different mindsets and these are rarely done equally well by the same mathematician. But no one doubts that algebraists and analysts are both mathematicians, even if they cannot get beyond the first page of each others' papers. Knuth [35] did an interesting study on the connections between algorithmic and mathematical thinking, a topic that he also touched on in an earlier paper [34] on "computer science and its relation to mathematics". This paper was written in 1974, at a time when many computer science departments were just being established, often by mathematicians. It makes interesting reading today, along with similar papers of the same vintage, such as [9], [25]. Software development is a logical conclusion of algorithmic thinking, and the development of software for mathematical algorithms naturally belongs in a mathematics department.

The reason I care about this topic is not for my own mathematical ego. I have been lucky to be at an institution where my work in this direction has been encouraged, or at least tolerated. It may have helped that I did not put much effort into software development until well after I was tenured.[1] The main value of the tenure system is that established people do not need to worry what our colleagues think of our activities or how they choose to label them. But the future depends on bright young people. I think computational science affords a wonderful opportunity to get students involved in a host of mathematical challenges, and making significant progress on these requires computational mathematicians with solid training in a broad range of mathematical

[1] However, many of my attitudes towards software development were shaped by my experiences as a graduate student in the Computer Science Department at Stanford, where students in the numerical analysis group were responsible for maintaining the library of numerical routines [10] available to physicists at the Stanford Linear Accelerator Center (SLAC) and acting as consultants, activities that were encouraged by Gene Golub and Joe Oliger. There I had the pleasure of working directly with an outstanding set of fellow students, most of whom have gone on to make software contributions of their own, including Marsha Berger, Petter Bjorstad, Dan Boley, Tony Chan, Bill Coughran, Bill Gropp, Eric Grosse, Mike Heath, Frank Luk, Stephen Nash, Michael Overton, and Lloyd Trefethen. Many of us were also shaped by Cleve Moler's course on numerical linear algebra, where he tried out his new MATLAB program on us as a front end to the LINPACK and EISPACK routines that were already setting the standard for mathematical software [23], [27], [61]. I think the Computer Science students were more impressed with MATLAB and much more influenced by this experience than Cleve takes credit for in [48].

tools and the ability to apply mathematical abstraction to common problems arising in multiple fields. While there are many talented computational scientists working on specific challenging problems in virtually every science and engineering department, a computational mathematician, centered in a mathematics (or applied mathematics) department, has the best chance of appreciating the common mathematical structure of the problems and producing algorithms and software that are broadly applicable. Doing so not only avoids a lot of wasted effort by scientists whose time is better spent on the peculiarities of their specialty, it also leads to the introduction of techniques into fields where they might not otherwise be invented and the discovery of new connections between existing algorithms and applications.

I once heard Jim Glimm remark that "applied mathematicians are the honey bees of the mathematical world, whose job is to cross-pollinate applications." In addition to providing a service to those in other disciplines, the process of collecting nectar can result in some sweet rewards back in our own hive. This is equally true in algorithm and software development as it is in more classical and theoretical aspects of applied mathematics.

But young mathematicians will feel free to pursue such activities, and to also do the less rewarding but crucial aspects of the scholarship such as documenting their codes and making them presentable to the rest of the world, only if it is accepted as valid mathematical scholarship. If it is seen as non-mathematics, it will only be a waste of time that is best avoided by anyone seeking tenure in a mathematics department.

Applied mathematics in general is becoming much more acceptable in mathematics departments than it once was, at least in the United States. However, I doubt that the careful development of software or computer programs, or the work required to turn research codes into publishable scholarship, has the same level of acceptance.

## 2. Numerical analysis, scientific computing, and computational science & engineering

One can argue at length about the meaning of the terms in this section title. To me, "numerical analysis" has a double meaning: the analysis and solution of real-world problems using numerical methods, and the invention and analysis of the methods themselves using the techniques of mathematics.

When used in the latter sense, numerical analysis belongs firmly in a mathematics (or applied mathematics) department. As just one example, analyzing the stability and convergence properties of finite difference or finite element methods is no less difficult (often more difficult) than analyzing the underlying differential equations, and relies on similar tools of analysis. Specialists in this type of numerical analysis may or may not do much computing themselves, and may be far removed from computational science.

Numerical analysis in the sense of using numerical methods to solve problems, or developing software for general use, is often called "scientific computing" or "com-

putational science & engineering" these days. One can make a further distinction between these two terms: Scientific computing is often used to refer to the development of computational tools useful in science and engineering. This is the main thrust of the *SIAM Journal on Scientific Computing*, for example, which contains few theorems relative to the more theoretical *SIAM Journal on Numerical Analysis*, but still focuses on mathematical and algorithmic developments. Computational science & engineering refers more specifically to the use of computational tools to do real science or engineering in some other field, as a complement to experimental or theoretical science and engineering.

Not everyone would agree with my definitions of these terms. In particular, it can be argued that "computational science" refers to the science of doing computation and "computational engineering" to the implementation of this science in the form of software development, but for my purposes I will lump these two activities under "scientific computing". It is important to be aware of this lack of consistency in nomenclature since, for example, many recently developed academic programs in Computational Science & Engineering stress aspects of scientific computing as well.

I have been arguing that "scientific computing", in the sense just described, is a branch of mathematics (as well as being a branch of other disciplines, such as computer science), and that other mathematicians should be more aware of the intellectual challenges and demands of this field, including the need to document and distribute code. Not only are the activities of many practitioners of scientific computing essentially mathematical, but they (and their students) benefit greatly from frequent contact with more theoretical numerical analysts and mathematicians working in related areas. Other mathematicians may also benefit from having computational experts in the department, particularly as more fields of pure mathematics develop computational sides and realize the benefits of experimental mathematics – there is even a journal (see expmath.org) now devoted to this approach.

## 3. Reproducible research

Within the world of science, computation is now rightly seen as a third vertex of a triangle complementing experiment and theory. However, as it is now often practiced, one can make a good case that computing is the last refuge of the scientific scoundrel. Of course not all computational scientists are scoundrels, any more than all patriots are, but those inclined to be sloppy in their work currently find themselves too much at home in the computational sciences. Buckheit and Donoho [13] refer to the situation in the field of wavelets as "a scandal". The same can be said of many other fields, and I include some of my own work in the category of scandalous.

Where else in science can one get away with publishing observations that are claimed to prove a theory or illustrate the success of a technique without having to give a careful description of the methods used, in sufficient detail that others can attempt to repeat the experiment? In other branches of science it is not only expected

that publications contain such details, it is also standard practice for other labs to attempt to repeat important experiments soon after they are published. Even though this may not lead to significant new publications, it is viewed as a valuable piece of scholarship and a necessary aspect of the scientific method.

Scientific and mathematical journals are filled with pretty pictures these days of computational experiments that the reader has no hope of repeating. Even brilliant and well intentioned computational scientists often do a poor job of presenting their work in a reproducible manner. The methods are often very vaguely defined, and even if they are carefully defined they would normally have to be implemented from scratch by the reader in order to test them. Most modern algorithms are so complicated that there is little hope of doing this properly. Many computer codes have evolved over time to the point where even the person running them and publishing the results knows little about some of the choices made in the implementation. And such poor records are typically kept of exactly what version of the code was used and the parameter values chosen that even the author of a paper often finds it impossible to reproduce the published results at a later time.

The idea of "reproducible research" in scientific computing is to archive and make publicly available all of the codes used to create the figures or tables in a paper in such a way that the reader can download the codes and run them to reproduce the results. The program can then be examined to see exactly what has been done.

The development of very high level programming languages has made it easier to share codes and generate reproducible research. Historically, many papers and text books contained pseudo-code, a high level description of an algorithm that is intended to clearly explain how it works, but that would not run directly on a computer. These days many algorithms can be written in languages such as MATLAB in a way that is both easy for the reader to comprehend and also executable, with all details intact. For example, we make heavy use of this in the recent paper [15]. We present various grid mappings that define logically rectangular grids in smooth domains without corners, such as those shown later in Figure 1, and all of the MATLAB codes needed to describe the various mappings are short enough to fit naturally in the paper. The associated webpage contains the longer CLAWPACK codes used to solve various hyperbolic test problems on these grids.

Trefethen's book on spectral methods [62] is a good example of a textbook along these lines, in which each figure is generated by a 1-page MATLAB program. These are all included in the book and nicely complement the mathematical description the methods discussed. Trefethen makes a plea for more attention to short and elegant computer programs in his recent essay [64].

However, for larger scale computer programs used in scientific publications, there are many possible objections to making them available in the form required to reproduce the research. I will discuss two of these, perhaps the primary stumbling blocks.

One natural objection is that it is a lot of work to clean up a code to the point where someone else can even use it, let alone read it. It certainly is, but it is often well

worth doing, not only in the interest of good science but also for the personal reason of being able to figure out later what you did and perhaps build on it.

Those of us in academia should get in the habit of teaching good programming, documentation, and record keeping practices to our students, and then demand it of them. We owe it to them to teach this set of computational science skills, ones that I hope will be increasingly necessary in academic research environments and that are also highly valued in industrial and government labs. It will also improve the chances that we will be able to build on the work they have done once they graduate, and that future students will be able to make use of it rather than starting from scratch as is too often the case today.

While ideally all published programs would be nicely structured and easily readable with ample comments, as a first step it would be valuable simply to provide and archive the working code that produced the results in a paper. Even this takes more effort than one might think. It is important to begin expecting this as a natural part of the process so that people will feel less like they have to make a choice between finishing off one project properly or going on to another where they can more rapidly produce additional publications. The current system strongly encourages the latter.

As Buckheit and Donoho [13] point out, the scientific method and style of presenting experiments in publications that is currently taken for granted in the experimental sciences was unheard of before the mid-1800s. Now it is a required aspect of respectable research and experimentalists are expected to spend a fair amount of time keeping careful lab books, fully documenting each experiment, and writing their papers to include the details needed to repeat the experiments. A paradigm shift of the same nature may be needed in the computational sciences.

Requiring it of our students may be a good place to start, provided we recognize how much time and effort it takes. Perhaps we should be more willing to accept an elegant and well documented computer program as a substantial part of a thesis, for example.

A second objection to publishing computer code is that a working program for solving a scientific or engineering problem is a valuable piece of intellectual property and there is no way to control its use by others once it is made publicly available. Of course if the research goal is to develop general software then it is desirable to have as many people using it as possible. However, for a scientist or mathematician who is primarily interested in studying some specific class of problems and has developed a computer program as a tool for that purpose, there is little incentive to give this tool away free to other researchers. This is particularly true if the program has taken years to develop and provides a competitive edge that could potentially lead to several additional publications in the future. By making the program globally available once the first publication appears, other researchers can potentially skip years of work and start applying the program to other problems immediately. In this sense providing a program is fundamentally different than carefully describing the materials and techniques of an experiment; it is more like inviting every scientist in the world to come use your carefully constructed lab apparatus free of charge.

This argument undoubtedly has considerable merit in some situations, but on the whole I think it is overblown. It is notoriously difficult to take someone else's code and apply it to a slightly different problem. This is true even when people are trying to collaborate and willing to provide hands-on assistance with the code (though of course this type of collaboration does frequently occur). It is often true even when the author of the code claims it is general software that is easy to adapt to new problems. It is particularly true if the code is obscurely written with few comments and the author is not willing to help out, as would probably be true of many of the research codes people feel the strongest attachment to.

Moreover, my own experience in computational science is that virtually every computational experiment leads to more questions than answers. There is such a wealth of interesting phenomena that can be explored computationally these days that any worthwhile code can probably lead to more publications than its author can possibly produce. If other researchers are able to take the code and apply it in some direction that would not otherwise be pursued, that should be seen as a positive development, both for science and for its original author, provided of course that s/he gets some credit in the process. This is particularly true for computational mathematicians, whose goals are often the development of a new algorithm rather than the solution of specific scientific problems. Even for those not interested in software development *per se*, anything we can do to make it easier for others to use the methods we invent should be viewed as beneficial to our own careers.

Perhaps what is needed is some sort of recognized patent process for scientific codes, so that programs could be made available for inspection and independent execution to verify results, but with the understanding that they cannot be modified and used in new publications without the express permission of the author for some period of years. Permission could be granted in return for co-authorship, for example. In fact such a system already works quite well informally, and greater emphasis on reproducible research would make it function even better. It would be quite easy to determine when people are violating this code of ethics if everyone were expected to "publish" their code along with any paper. If the code is an unauthorized modification of someone else's, this would be hard to hide.

## 4. Wave propagation algorithms and CLAWPACK

As a case study in software development, and its relation to mathematics, scientific computing, and reproducible research, I will briefly review some of the history behind my own work on CLAWPACK (Conservation LAWs PACKage), software for solving hyperbolic systems of partial differential equations.

This software development project began in 1994. I had just taught a graduate course on numerical methods for conservation laws and had distributed some sample computer programs to the students as a starting point for a class project. In the fall I went on sabbatical and decided to spend a few weeks cleaning up the program and

redesigning it as a software package, in large part because I was also planning to spend much of the year revising my lecture notes [42] from a course I taught at ETH-Zürich in 1989 into a longer book, and I wanted to complement the text with programs the students could easily use.

I seriously misjudged the effort involved – I spent most of that year and considerable time since developing software, which grew into something much more that I originally intended. The book [44] took several more years of work and iterations of teaching the course and did not appear until 2002.

Virtually all of the figures in this book are reproducible, in the sense that the programs that generated them can each be downloaded from a website and easily run by the student. Most figure captions contain a link to the corresponding website, in the form [claw/book/chap23/advection/polar], for example, from the caption of Figure 23.3, which is easily translated into the appropriate web address. (Start at http://www.amath.washington.edu/~claw/book.html to browse through all these webpages). Each webpage contains the computer code and many also contain additional material not in the book, for example movies of the solution evolving in time.

All of these examples are based on the CLAWPACK software. This software is described briefly in the book and more completely in the User Guide [41] available on the web. Once the basic software is installed, the problem-specific code for each example in the book is quite small and easy to comprehend and modify. The reader is encouraged to experiment with the programs and observe how changes in parameters or methods affect the results. These programs, along with others on the CLAWPACK website [40], can also form the basis for developing programs to solve similar problems.

This software implements a class of methods I call "wave propagation algorithms" for solving linear or nonlinear hyperbolic problems. Hyperbolic partial differential equations are a broad class of equations that typically model wave propagation or advective transport phenomena. The classic example is the second-order wave equation $p_{tt} = c^2 p_{xx}$ for linear acoustics, modeling the propagation of pressure disturbances in a medium with sound speed $c$. The CLAWPACK software, however, is set up to solve a different form of hyperbolic equations: systems that involve only first order derivatives in space and time.

In the linear case, a first-order system of PDEs (in one space dimension and time) has the form $q_t + Aq_x = 0$, where $q(x, t)$ is a vector of some $m$ conserved quantities, $A$ is an $m \times m$ matrix, and subscripts denote partial derivatives. In the nonlinear case, a system of $m$ conservation laws takes the form $q_t + f(q)_x = 0$, where $f(q)$ is the flux function (in the linear case, $f(q) = Aq$). This system is called *hyperbolic* if the flux Jacobian matrix $f'(q)$ is diagonalizable with real eigenvalues. The system of Euler equations for inviscid compressible gas dynamics has this form, for example, where mass, momentum, and energy are the conserved quantities. The full nonlinear equations can develop shock wave solutions in which these quantities are discontinuous, one of the primary challenges in numerical modeling. Linearizing this system gives the linear acoustics equations in the form of a first-order system of

equations for pressure and velocity. Cross differentiating this system allows one to eliminate velocity and obtain the single second-order wave equation for $p$ mentioned above, but the first-order formulation allows the modeling of a much broader range of phenomena.

First-order hyperbolic systems arise naturally in a multitude of applications, including for example elastodynamics (linear and nonlinear), electromagnetic wave propagation (including nonlinear optics), shallow water equations (important in oceanography and atmospheric modeling), and magnetohydrodynamic and relativistic flow problems in astrophysics.

The wave-propagation algorithms are based on two key ideas: Riemann solvers and limiters. The *Riemann problem* consists of the hyperbolic equation under study with special initial conditions at some time $\bar{t}$: piecewise constant data with left state $q_\ell$ and right state $q_r$ and a jump discontinuity in each conserved quantity at a single point in space, say $\bar{x}$. The solution to this Riemann problem for $t > \bar{t}$ is a similarity solution, a function of $(x-\bar{x})/(t-\bar{t})$ alone that consists of a set of waves propagating at constant speeds away from the initial discontinuity. The definition of hyperbolicity guarantees this, and the eigenvalues of the flux Jacobian are related to the wave speeds. In the linear case the eigenvalues of the matrix $A$ are exactly the wave speeds. In the nonlinear case the eigenvalues vary with $q$. The Riemann solution may then contain shock waves and rarefaction waves, but even in the nonlinear case this special problem has a similarity solution with constant wave speeds.

In 1959, Sergei Godunov [30] proposed a numerical method for solving general shock wave problems in gas dynamics by using the Riemann problem as a building block. If the physical domain is decomposed into a finite number of grid cells and the solution approximated by a piecewise constant function that is constant in each grid cell, then at the start of a time step the initial data consists of a set of Riemann problems, one at each cell interface. By solving each of these Riemann problems the solution can be evolved forward in time by a small increment. The resulting solution is averaged over each grid cell to obtain a new piecewise constant approximation to the solution at the end of the time step. This procedure is repeated in the next time step. This idea of basing the numerical method on Riemann solutions turned out to be a key idea in extending the "method of characteristics" from linear hyperbolic systems to important nonlinear shock propagation problems. This also leads naturally to a software framework: the particular hyperbolic equation being solved is determined by providing a Riemann Solver. This is a subroutine that, given any two states $q_\ell$ and $q_r$, returns the wave speeds of the resulting waves in the similarity solution, along with the corresponding waves themselves (i.e., the jump in $q$ across each wave). The updating formulas for the cell averages based on these waves are very simple and independent of the particular system being solved.

Godunov's method turned out to be very robust and capable of solving problems involving strong shocks where other methods failed. However, it is only first-order accurate on smooth solutions to the PDEs. This means that the error goes to zero only as the first power of the discretization steps $\Delta x$ and $\Delta t$. Moreover, although

complicated solutions involving strong shocks and their interactions could be robustly approximated without the code crashing, the resulting approximations of shock waves are typically smeared out. The process of averaging the solution over grid cells each time step introduces a large amount of "numerical dissipation" or "numerical viscosity".

During the 1970s and 1980s, a tremendous amount of effort was devoted to developing more accurate versions of Godunov's method that better approximated smooth solutions and also captured shock waves more sharply. These methods often go by the general name of "high-resolution shock capturing methods". A wide variety of methods of this type have been proposed and effectively used. Many of these, including the wave-propagation algorithms of CLAWPACK, have the relatively modest goal of achieving something close to second-order accuracy on smooth solutions coupled with sharp resolution of discontinuities. Other approaches have been used that can achieve much better accuracy in certain situations, though for general nonlinear problems involving complicated shock structures, particularly in more than one dimension, it seems hard to improve very much beyond what is obtained using second-order methods.

One standard second-order method for a linear hyperbolic system is the Lax–Wendroff method, first proposed in 1960, which is based on approximating the first few terms of a Taylor series expansion of the solution at time $t + \Delta t$ about the solution at time $t$. This method does not work at all well for problems with discontinuities, however, as it is highly dispersive and nonphysical oscillations arise that destroy all accuracy. In the nonlinear case these oscillations around shock waves can also lead to nonlinear instabilities.

The key feature in many high-resolution methods is to apply a *limiter function* in some manner to suppress these oscillations. In the wave-propagation algorithms this is done in the following way. The Lax–Wendroff method can be rewritten as Godunov's method plus a correction term that again can be expressed solely in terms of the waves and wave speeds in the Riemann solutions arising at each cell interface. Where the solution is smooth, adding in these correction terms improves the accuracy. Where the solution is not smooth, for example near a shock, the Taylor series expansion is not valid and these "correction terms", which approximate higher derivatives in the Taylor expansions, do more harm than good. We can determine how smooth the solution is by comparing the magnitude of a wave with the magnitude of the corresponding wave at the neighboring cell interfaces. If these differ greatly then the solution is not behaving smoothly and the correction terms should be "limited" in some manner.

Many variants of this idea have been used. In some cases it is the Lax–Wendroff expression for the flux at the interface between cells that is limited (in so-called "flux limiter" methods). Another approach is to view the Lax–Wendroff method as a generalization of Godunov's method in which a piecewise linear function in each grid cell is defined and the values at the edges of each cell then used to define Riemann problems. In this case the slope chosen in each cell is based on the averages in nearby cells, with some "slope limiter" applied in regions where it appears the solution is not behaving smoothly. For the special case of a scalar conservation law, a very

nice mathematical theory was developed to guide the choice of limiter functions. The true solution to a scalar problem has its total variation non-increasing over time. By requiring that a numerical method be "total variation diminishing" (TVD) it was possible to derive methods that were essentially second order accurate, or higher, but that could be proved to not introduce spurious oscillations.

The 1980s were an exciting time in this field, as robust high-resolution methods were developed for a variety of challenging applications and as computers became powerful enough that extensions of these methods and the related mathematical theory to more than one space dimension became possible and necessary.

I played a modest role in some of these developments, but I think my main contribution has been in providing a formulation of these methods that lends itself well to software that is very broadly applicable, and in leading the effort to write this software. The wave-propagation formulation that I favor has the advantage that once the Riemann problem has been solved, the limiters and high-resolution correction terms are applied in a general manner that is independent of the equation being solved. Moreover a framework for doing this in two dimensions was proposed in [43] that retains this modularity, separating the process of solving a one-dimensional Riemann problem at each cell interface, along with a related "transverse Riemann problem" in the orthogonal direction, from the process of propagating these waves and correction terms with appropriate limiters. While it had long been recognized that the methods being developed were in principle broadly applicable to all hyperbolic problems, for systems of equations most methods were developed or at least presented in a form that was specific to one particular problem (often the Euler equations of gas dynamics) and a certain amount of work was required to translate them to other problems. Computer programs that I was aware of were all problem-specific, and generally not publicly available.

My original motivation for developing this framework was not software development, but rather the need to teach graduate students, and the desire to write a book that explained how to apply these powerful high-resolution methods to a variety of problems. I had students coming from the Mathematics and Applied Mathematics departments, many of whom knew little about fluid dynamics, and students from several science and engineering departments who had specific interests in very diverse directions.

My subsequent motivation for software development was partly educational, but also partly because I wanted to better publicize the wave-propagation framework I had developed and make it easier for others to use methods in this form. I recognized that these methods, while quite general, were also sufficiently complicated that few would bother to implement them from my descriptions in journal articles. For this research to have any impact beyond a few publications it seemed necessary to provide more than pseudo-code descriptions of the algorithms.

## 5. CLAWPACK as an environment for developing and testing methods

I have provided a few basic sample programs within the CLAWPACK package itself, and some additional test problems in a directory tree labeled applications available at the website. Complete programs also exist for each of the figures in my book. These each consist of a tarred Unix directory containing a small set of subroutines to be used together with the main software. These subroutines specify the specific problem, including the Riemann solver and the initial and boundary conditions. Many standard boundary conditions are already available in the default boundary condition routine, and a variety of Riemann solvers are also provided for many different systems of equations. As a result it is often very easy to adapt one of these examples to a new problem, particularly for some of the standard test problems that often appear in publications.

Although the development of the CLAWPACK software was originally motivated by the desire to make a set of existing methods more broadly accessible, the availability of this software has also encouraged me to pursue new algorithmic advances that I otherwise might not have. I hope that the software will also prove useful to others as a programming environment for developing and testing new algorithms, and for comparing different methods on the same problems. Since the source code is available and the basic CLAWPACK routines are reasonably simple and well documented, it should be easy for users to modify them and try out new ideas. I encourage such use, and I certainly use it this way myself.

As one example, there are many approaches to developing the "approximate Riemann solvers" that are typically used for nonlinear problems. It often is not cost effective, and may not even be possible, to solve the Riemann problem exactly at every cell interface each time step. Different versions of the Riemann solver can easily be swapped in and tested on a set of problems. See Figures 15.5 and 15.6 in [44] and the associated programs for an example of this sort of comparison. The CLAWPACK software also comes with a set of standard limiter functions, and an input parameter specifies which one will be used. The subroutine where these are implemented can be easily modified to test out new approaches.

More extensive modifications could be made to the software as well, for example by replacing the wave-propagation algorithms currently implemented by a different approach. If a new method is formulated so that it depends on a Riemann solver and boundary conditions that are specified in the form already used in CLAWPACK, then it should be easy to test out the new method on all the test problems and applications already developed for CLAWPACK. This would facilitate comparison of a new method with existing methods.

Careful direct comparisons of different methods on the same test problems are too seldom performed in this field, as in many computational fields. One reason for this is the difficulty of implementing other peoples' methods, so the typical paper contains only results obtained with the authors' method. Sometimes (not always) the method has been tested on standard test problems and the results can be compared with

others in the literature, with some work on the reader's part, and assuming the reader is content with comparisons in the "eyeball norm" since many papers only contain contour plots of the computed solution and no quantitative results. Of course there are many exceptions to this, including some papers devoted to careful comparisons of different methods (e.g., [31], [46]), but these papers are still a minority. I hope that CLAWPACK might facilitate this process more in the future, and that new algorithms of this same general type might be provided in a CLAWPACK implementation that allows direct use and comparison by others.

## 6. CLAWPACK extensions and infrastructure

This software effort began about 10 years ago, as recounted above, and has continued in unexpected ways as the software grew beyond my initial intentions in several directions. While originally I viewed it primarily as an educational tool, it was based on my own research codes and I realized that it could perhaps be valuable for other researchers as well, perhaps even as a tool for solving real problems and not just academic test problems. To make it more useful as a general tool required several enhancements.

I originally wrote subroutines for solving systems in one and two space dimensions, but started collaborating with Jan Olav Langseth that same year on the development of three-dimensional generalizations [37]. He took my course as a visiting graduate student from the University of Oslo and went on to write a thesis partly on this topic [36], and wrote the three-dimensional subroutines in CLAWPACK.

I had also been collaborating with Marsha Berger on research related to using cut-cell Cartesian grids to solve the Euler equations in non-rectangular geometries (e.g., [5], [6]) and was familiar with her implementation of adaptive mesh refinement on rectangular grids, using an approach pioneered in her work with Joe Oliger and Phil Colella [4], [8], [7]. While still on sabbatical, in 1995 we started working together to modify her adaptive mesh refinement (AMR) code and make it more generally applicable to any hyperbolic system that can be solved in CLAWPACK. In this approach the rectangular grid is refined by introducing finer grids on rectangular patches. Several levels of refinement are allowed, by an arbitrary refinement factor at each level. The program automatically refines and de-refines as the computation proceeds, based on a standard default or user-specified error criterion, and so fine grids are used only where they are needed. This approach is particularly valuable for problems involving shock waves propagating in three space dimensions. To capture the shock waves as sharp discontinuities, even with high-resolution methods, requires a fairly fine grid around the shock, much finer than is needed in regions where the solution is smooth. Without AMR many three-dimensional problems would be impractical to solve except on the largest supercomputers.

This AMRCLAW software is now a standard part of CLAWPACK, and was extended from two to three space dimensions with programming help from Dave McQueen and

Donna Calhoun. The goal of the AMRCLAW code is primarily to facilitate research and practical problem solving, rather than to teach adaptive refinement techniques. The core library consists of about 9 000 lines of Fortran 77 code that is quite convoluted and not structured or fully documented in a way that others can easily understand. This stems in large part from its history as a merging together of two different codes, with different notation and conventions that have largely been preserved and worked around, and with many new features added over the years as afterthoughts that were not originally designed for. It may not be particularly elegant, but it has proved valuable in solving practical problems and by now has been tested to the point where it is fairly robust and reliable. All of the source code is available for others to inspect and modify, at their own peril perhaps.

The CLAWPACK framework can also be used with other AMR packages. Sorin Mitran is developing BEARCLAW [47], a Fortran 90 version with similar capabilities to AMRCLAW but designed with these capabilities in mind from the beginning. This code is still being developed and has not been tested as extensively as AMRCLAW, but it has been successfully used in astrophysical applications [51], [52]. Ralf Deiterding has developed a general purpose adaptive refinement code AMROC [18], [19] in C++ that also allows the use of the CLAWPACK solvers with adaptive refinement. This is a primary computational tool in the Virtual Shock Physics Test Facility at the Caltech ASC Center for Simulation of Dynamic Response of Materials [1]. Recently Donna Calhoun has written a CHOMBO-CLAW interface [14] between the CLAWPACK solvers and the C++ CHOMBO software package developed by Colella's group at Lawrence Berkeley Lab [16]. These newer packages have various advantages over the AMRCLAW software in CLAWPACK. They are written in more advanced languages that are more suitable for the data structures and dynamic memory allocation needed in AMR codes. They also have more capabilities such as parallel implementations, the ability to handle implicit methods, and/or coupling with elliptic solvers as required in some applications.

Other extensions of CLAWPACK have also been developed, such as the CLAWMAN software [2] that solves hyperbolic systems on curved two-dimensional manifolds. This has been used to solve geophysical flow problems on the sphere and relativistic flow problems in curved space-time near a black hole [3], [54], [55].

The original software was designed for purely Cartesian grids in rectangular regions of space. Some practical problems have this form, but most are posed in more complicated physical domains, e.g., for flows around or through a physical object. There are many approaches to handling complex geometries. The cut-cell Cartesian grid approach has already been mentioned above. At the other extreme lie unstructured grids, typically composed of triangular cells in two dimensions or tetrahedra in three dimensions. These can conform to very general boundaries, but grid generation then becomes a challenging problem in itself, and implementations must deal with special data structures to keep track of what cells are adjacent to one another.

For fairly simple domains, a good compromise is often possible in which the grid is logically rectangular in computational space, but is mapped to a nonrectagular

physical domain. In two dimensions, this means that each grid cell is a quadrilateral and simple $(i, j)$ indexing can be used to denote the cells, with neighboring cells having indices $(i \pm 1, j \pm 1)$. In three dimensions the grid cells are hexahedral but still logically rectangular. It is quite easy to apply CLAWPACK in such situations (as described in Chapter 23 of [44]), with a standard set of additional subroutines used to specify the mapping function. One nice feature of this approach is that the AMR routines work perfectly well on mapped grids – the patches of refinement are still rectangular in computational space. The wave-propagation algorithms turn out to work quite robustly on quadrilateral grids, even if the grid is nonorthogonal and far from smooth. Rather than incorporating the grid mapping directly into the differential equations as "metric terms" that involve derivatives of the mapping function, in the wave-propagation approach one solves one-dimensional Riemann problems orthogonal to each grid interface and transverse Riemann problems based on the adjacent cell interfaces.

Figure 1 shows two grids from some recent work with Calhoun [15] on the use of logically rectangular grids for solving problems in domains with smooth boundaries. The figure on the left shows a quadrilateral grid for a circle, while that on the right is a logically rectangular grid on the sphere. Each grid is simply a rectangle in computational space. Of course polar coordinates also give a logically rectangular grid in a circle, but grid lines coalesce at the center where cells have much smaller area than those near the perimeter. This presents a problem when using explicit

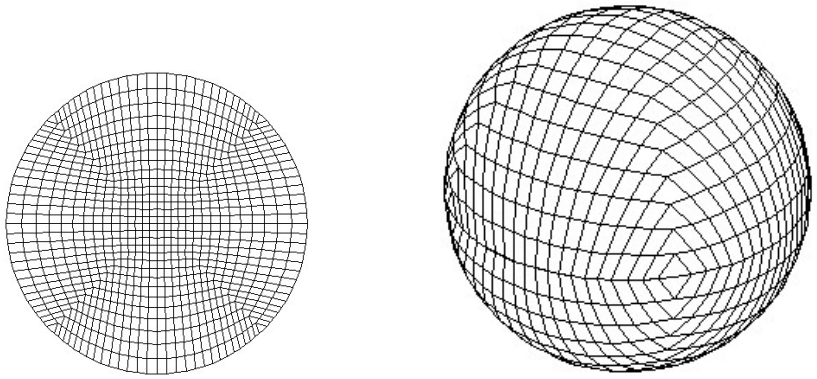


Figure 1. Quadrilateral grids on the circle and the sphere. In each case the computational domain is a rectangular grid. The MATLAB code that generates these grids is available on the webpage [39].

methods for hyperbolic problems: the disparity in cell sizes leads to an undesirable restriction on the time step. The grids in Figure 1 are far from smooth or orthogonal; in fact the images of the two orthogonal cell edges at each corner of the rectangular computational domain are nearly collinear in the physical domain. But the cell areas are nearly uniform, differing by at most a factor of 2. Standard finite difference

methods would presumably not work well on these grids, but applying CLAWPACK yields very nice results, as demonstrated in [15].

Writing new methods in "CLAWPACK form" makes it possible to take advantage of the infrastructure developed for this software, for example to apply the method on a general quadrilateral grid. It may also be possible to apply adaptive mesh refinement quite easily with the new method by taking advantage of one of the AMR wrappers described above. Since implementing AMR effectively is far from trivial, and often has little to do with the particular numerical method used on each grid patch, making use of existing implementations could be worthwhile even if it takes a bit of work to formulate the method in an appropriate form. Extensive graphics routines in MATLAB have been written to plot the results computed by CLAWPACK. These routines (written mostly by Calhoun and myself) deal with adaptive mesh refinement data in two or three space dimensions. This is a not easily done directly with most graphics packages, and again this infrastructure may be useful to developers of new methods. (In fact these graphics routines can take AMR data produced by any program, not just CLAWPACK, provided it is stored in the appropriate form.)

One current research project, with graduate student David Ketcheson, is to implement higher order numerical methods in CLAWPACK, specifically the weighted essentially non-oscillatory (WENO) methods that are based on higher order interpolation in space coupled with Runge–Kutta time stepping (e.g., [31], [60]). Doing so requires reformulating these methods slightly, and in particular we are developing a version that works for linear hyperbolic systems that are not in conservation form. Systems of this form arise in many applications, such as the propagation in heterogeneous media of acoustic, elastic, or electromagnetic waves. In many linear applications the solution is smooth but highly oscillatory, and in this case higher-order methods may be beneficial.

## 7. Applications in computational science

In addition to continuing to work on algorithm development, in the past few years I have become more directly involved in applications of computational science, driven in part by the existence of this software as a starting point. Problems that can be solved easily with the existing software have little interest to me; as a mathematician I consider them solved, though there may be plenty of interesting science to be done by judicious use of the software as a tool. This is best done, however, by scientists who are experts in a particular domain.

A practical problem where CLAWPACK fails to perform well, or where some substantial work is required to apply it, is much more interesting to me. Although the methods implemented in CLAWPACK work well on many classical hyperbolic problems, there are a wealth of more challenging problems that are yet to be solved, and I hope that CLAWPACK might form the basis for approaching some of these problems without the need to rewrite much of the basic infrastructure.

In the remainder of this section I will briefly describe a few topics that are currently occupying me and my students. More details on these and other problems, along with movies, papers, and sometimes computer code, can be found by clicking on the "Research interests" link from my webpage.

**Shock wave therapy and lithotripsy.** Focused shock waves are used in several medical procedures. Extracorporeal shock wave lithotripsy (ESWL) is a standard clinical procedure for pulverizing kidney stones noninvasively. There are several different lithotripter designs. In one model, a spark plug immersed in water generates a cavitating bubble that launches a spherical shock wave. This wave reflects from an ellipsoidal shaped reflector and refocuses at the distal focus of the ellipsoid, where the kidney stone is centered. The shock wave pulse has a jump in pressure of roughly 50 MPa (500 atmospheres) over a few nanoseconds, followed by a more slowly decaying decrease in pressure passing below atmospheric pressure before relaxing to ambient. This tensile portion of the wave is particularly important in kidney stone comminution since stones are composed of brittle material that does not withstand tensile stress well. Typically thousands of pulses are applied clinically (at a rate of 1 to 4 per second). The breakup process is not well understood and better understanding might allow clinical treatment with fewer pulses and less damage to the surrounding kidney.

Together with Kirsten Fagnan and Brian MacConaghy, two graduate students in Applied Mathematics, I have recently been collaborating with researchers at the Applied Physics Laboratory and the medical school at the University of Washington to develop a computational model of nonlinear elastic wave propagation in heterogeneous media that can be used to aid in the study of this process. Preliminary computations have been performed using linear elasticity in two-dimensional axisymmetric configurations in which a cylindrical test stone is aligned with the axis of the lithotripter ellipsoid. We are currently extending these computations to an appropriate nonlinear elasticity model, and also to three dimensional calculations for non-axisymmetric configurations.

We also hope to perform simulations useful in the study of extracorporeal shock wave therapy (ESWT), a relatively new application of lithotripter shock waves to treat medical conditions other than kidney stones, in which the goal is to stimulate tissue or bone without destroying it. For example, several recent clinical studies have shown that treating nonunions (broken bones that fail to heal) with ESWT can lead to rapid healing of the bone [26], [57], perhaps because it stimulates the growth of new vascular structure in regions where there is insufficient blood flow. Conditions such as tennis elbow, plantar fasciitis, and tendinitis have also been successfully treated with ESWL; see for example [32], [58].

In ESWL applications the shock wave is often focusing in a region where there is a complicated mix of bones and tissue. The interfaces between these materials cause significant reflection of wave energy. It is often crucial to insure that the shocks do not accidentally refocus in undesirable locations, such as nearby organs or nerves, which could potentially cause extensive collateral damage. Ideally one would like to be able

to use MRI data from a patient to set material parameters and run 3d simulations of the shock wave propagation in order to adjust the angle of the beam to achieve maximal impact in the desired region with minimal focusing elsewhere. Our current work is a first step in this direction.

**Volcanic flows.** My recent student Marica Pelanti developed a dusty gas model in which the Euler equations for atmospheric gas are coupled to another set of conservation laws for the mass, momentum, and energy of a dispersed dust phase [49], [50]. The two sets of equations are coupled together by source terms modeling viscous drag and heat transfer between the phases. The dust is assumed pressureless and requires special Riemann solvers, based on [45]. This model has been used to study the jets arising from high-velocity volcanic eruptions.

The speed of sound in a dusty gas is considerably less than the sound speed in the atmosphere. As a result, volcanic jets can easily be supersonic relative to the dusty gas sound speed, leading to interesting shock wave structures within an eruption column. Pelanti explored this for 2D axi-symmetric jets for both flat topography and for topography where the jet expands through an idealized conical crater. Similar work has been performed by Augusto Neri and Tomaso Esposti Ongaro in the Earth Sciences Department at the University of Pisa and we are now collaborating with them on some comparisons of our results.

We have also interacted extensively with researchers at the USGS Cascade Volcano Observatory (CVO) near Mount St. Helens (MSH), who study many aspects of volcanic flows and are charged with hazard assessment for MSH. After a long quiescent period, MSH became quite active again in October, 2004, and we worked with these researchers to try to produce a full three-dimensional model of pyroclastic flows over the topography of MSH in order to predict the possible impact of an eruption of various magnitudes. Extension of the dusty gas model to the full three-dimensional topography of MSH is underway, although it turns out there are many numerical and modeling issues still to be tackled.

Considerable data is available from the 1980 eruption that can be used to validate a code. In particular, there is a set of photographs and maps that show the direction in which trees were blown down when the initial pyroclastic blast from the eruption passed over the surrounding ridges. The blown-down trees created a snapshot of the velocity vectors in the leading edge of the flow. These exhibit complex flow patterns, such as recirculation zones on the lee side of ridges where the trees were blown down in the direction pointing towards MSH instead of away. We hope to eventually compare computed velocities from our simulations with these observations.

Richard Iverson and Roger Denlinger at CVO have also done extensive work on modeling debris flows, such as those that arise when water from melting glaciers on a volcano mixes with trees, boulders, and other debris, creating highly damaging and life-threatening flows [20], [21]. Their numerical work is based on equations similar to the shallow water equations but enhanced with more physics within the flow. They use Riemann solver techniques based in part on the wave-propagation framework, with the

additional solution of an elasticity problem within each Riemann solver to compute the local stress tensor within the debris. They encounter difficulties at the edge of the flow similar to the dry-cell problems that arise in tsunami modeling (discussed below) and we have collaborated with them on solving these problems. Denlinger has also recently modeled the great Missoula floods using similar techniques [22].

**Tsunami modeling.** David George and I have been developing a version of AMRCLAW capable of modeling tsunamis, including both their global propagation over large expanses of ocean and the inundation and run-up around small scale features at the level of individual beaches or harbors. We solve the shallow water equations on rectangular grids, in which each grid cell has an elevation value for the earth surface (which is called topography if it is above sea level, or bathymetry when underwater). The components of $q$ are the fluid depth $h$ and momenta $hu$ and $hv$. Grid cells above sea level are dry ($h = 0$) and cells can become wet or dry dynamically as waves move along the shore. This approach avoids the need to model the shoreline as a separate interface, but developing a robust code based on this approach requires a Riemann solver that can deal well with both wet and dry states. The bathymetry comes in as a source term in the conservation laws. Away from shore the bathymetry is varying on a scale of several kilometers (the Indian Ocean is about 4 km deep, for example) whereas a tsunami propagating over the ocean is a few meters high at most. This difference in scales leads to difficulties that I will not describe here, and requires the use of some sort of "well-balanced" scheme in which the source term is incorporated into the Riemann solver. Though of small magnitude in the ocean, a tsunami may have a wavelength of more than 100 km and so the shallow water equations are an appropriate model. The propagation velocity is $\sqrt{gh}$, where $g$ is the gravitational constant, and as they approach shore $h$ decreases and the wave magnitude increases as the wavelength shortens, the same phenomenon observed in breaking waves on a beach. But in a tsunami the entire water column is set in motion by an uplifting of the ocean floor, whereas in wind-driven surface waves only the water very near the surface is moving. The enormous energy tsunamis carry gives them great destructive potential.

Adaptive mesh refinement is essential for this problem. We wish to propagate waves over the ocean, where grid cells several kilometers on a side can be used, and simultaneously predict the run-up around local features, where cell sizes of a few meters are desirable. Developing a well-balanced dry-state Riemann solver that works well in the context of AMR proved to be quite challenging and many difficulties appeared at the boundaries between grids at different levels. These could only be solved by some substantial reworking of the AMRCLAW code. The result is a special-purpose program that incorporates these algorithmic modifications and can now be applied to many tsunami scenarios. It is currently being tested by comparing predictions with measurements made at various places around the Indian Ocean in the wake of the 26 December 2004 Sumatra earthquake.

For this application we are working very closely with tsunami scientists. Our involvement in tsunami modeling arose out of a joint NSF grant with Harry Yeh in

civil engineering at Oregon State University and Joe Hamack at Penn State, who were doing wave tank experiments and related mathematical modeling that they wished to complement with numerical simulations. Since the Sumatra event, our focus has shifted to the larger scale, and the contacts we had already established in the tsunami modeling community proved invaluable. Many Tsunami Survey Teams traveled to the Indian Ocean and surveyed different parts of the coastline, measuring the run-up and inundation observed. Yeh was on a team that mapped the region near Chennai (Madras), India, and our initial validation work is focused on comparing predictions with his observations in this area. Unfortunately fine-scale bathymetry data is hard to come by and we have resorted to digitizing navigational charts to obtain some of the necessary data.

Figure 2 shows part of a simulation of the Indian Ocean tsunami, as described further in the caption. See the webpage [39] for color versions and movies, along with the computer program.

The program we have developed is a research code for this particular problem, but we intend to further improve it and ultimately make it available to the community. There is far more data available than we can compare against ourselves and we hope that other researchers will be able to use it for some of this work and publish the results. If our code does not work well or does not agree well with observations in some cases then we may need to revisit it, or perhaps others will make further improvements to it.

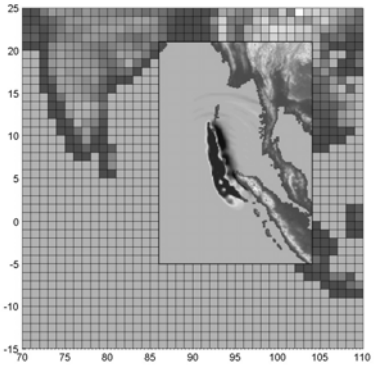
We hope that this code may ultimately be useful as a real-time tsunami prediction tool, and we are working with Vasily Titov and other scientists of the NOAA National Tsunami Hazard Mitigation Program in Seattle to compare our code with theirs and see how we can best complement their efforts (some of which are described in a recent Scientific American article [28]).

## 8. Conclusions

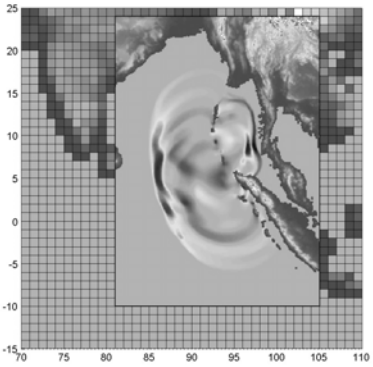
I have made a case that many aspects of scientific computing and software development should be viewed as inherently mathematical, and that mathematicians can play a very important role in computational science. I have also encouraged researchers in this area to produce reproducible research, in particular by making computer programs, not just software, available to others. I have presented some of my own research activities as a case study, though I do not claim it is the best example to follow.

I do hope, however, that the software we have produced will find wider use as one tool in this direction, both as a development environment for testing new methods and as a building block for solving problems in science and engineering. My hope is that others who develop methods or applications using this package will make their full code available on the web, particularly if it has been used to compute results that appear in publications.

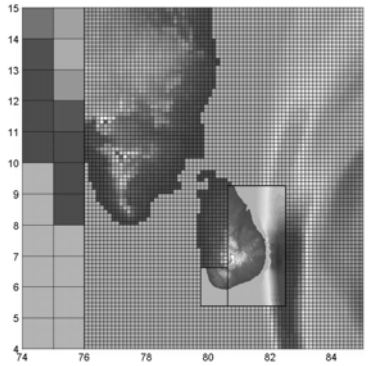
(a) Time 01:06:45 (525 seconds)




(b) Time 02:10:55 (4375 seconds)




(c) Time 02:43:00 (6300 seconds)




(d) Time 3:15:05 (8225 seconds)

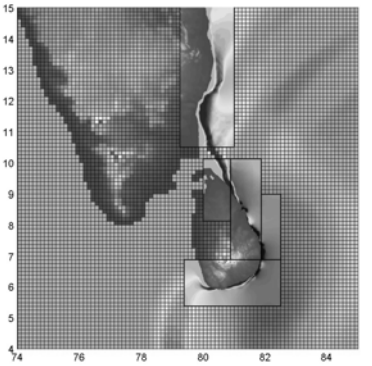



Figure 2. Propagation of the 26 December 2004 tsunami across the Indian Ocean. Units on the axes are longitude and latitude. The top two frames show the Bay of Bengal at two early times. A coarse grid is used where nothing is yet happening and the grid cells are shown on this "Level 1 grid", which has a mesh width of one degree (approximately 111 km). The rectangular region where no grid lines are shown is a Level 2 grid with mesh width 8 times smaller, about 14 km. Red represents water elevation above sea level, dark blue is below the undisturbed surface (see the webpage [39] for color versions of these images). Figure (c) shows a zoomed view of the southern tip of India and Sri Lanka at a later time. The original Level 1 grid is still visible along the left-most edge, but the rest of the region shown has been refined by a Level 2 grid. Part of Sri Lanka has been refined by Level 3 grids. The grid lines on Level 3 are not shown; the mesh width on this level is about 1.7 km, a factor of 8 finer than Level 2. Figure (d) shows a later time, as the wave diffracts around Sri Lanka, moving slowly through the shallow water in this coastal region. The calculation shown here was run on a single-processor 3.2 GHz PC under Linux and took about 40 minutes of wall time for the computation shown here. Movies of this simulation can be viewed on the webpage [39], which also contains pointers to finer grid calculations and more recent work on this problem.

Even for results that are not published, it would be valuable to have more examples and test problems available on-line than what is provided on the CLAWPACK web pages. Please let me know with a brief email if you have created such a page, or published a paper where CLAWPACK was successfully used.

I am also always interested to hear about problems that arise with the software or suggestions for improvements, though as an academic researcher with a small group of graduate students I cannot promise to provide as much technical support as I would like to.

The website [39] contains the codes used to generate the two figures in this paper, two very different examples of what can be provided in conjunction with a publication. The programs for Figure 1 are each less than a page of MATLAB, while the program for Figure 2 is about 13,000 lines of Fortran and also requires a large set of bathymetry and earthquake source data. The webpages also contain movies that illustrate the figures much better than the static versions shown in this paper, and links to other related work.

**Acknowledgments.** Numerous people read drafts of this paper and provided valuable comments and pointers to related work. In particular I would like to thank Marsha Berger, Donna Calhoun, Benjamin LeVeque, William LeVeque, and Nick Trefethen for their extensive comments, and David George for pulling together the necessary pieces for Figure 2.

## References

[1] *Caltech center for simulation of dynamic response of materials*. http://csdrm.caltech.edu/, 2005.

[2] Bale, D., Rossmanith, J. A., and LeVeque, R. J., CLAWMAN software. http://www.amath.washington.edu/~claw/clawman.html.

[3] Bale, D. S., Wave propagation algorithms on curved manifolds with applications to relativistic hydrodynamics. PhD thesis, University of Washington, 2002. http://www.amath.washington.edu/~rjl/people.html.

[4] Berger, M., Adaptive mesh refinement for hyperbolic partial differential equations. PhD thesis, Computer Science Department, Stanford University, 1982.

[5] Berger, M., and LeVeque, R. J., Cartesian meshes and adaptive mesh refinement for hyperbolic partial differential equations. In *Hyperbolic problems. Theory, numerical methods and applications* (ed. by B. Engquist and B. Gustafsson), Vol. I, Studentlitteratur, Lund 1991, 67–73.

[6] Berger, M., and LeVeque, R. J., Stable boundary conditions for Cartesian grid calculations. *Computing Systems in Engineering* **1** (1990), 305–311.

[7] Berger, M., and Oliger, J., Adaptive mesh refinement for hyperbolic partial differential equations. *J. Comput. Phys.* **53** (1984), 484–512.

[8] Berger, M., and Colella, P., Local adaptive mesh refinement for shock hydrodynamics. *J. Comput. Phys.* **82** (1989), 64–84.

[9] Birkhoff, G., Mathematics and computer science. *American Scientist* **63** (1975), 83–91.

[10] Bolstad, J. H., Chan, T. F., Coughran, W. M., Jr., Gropp, W. D., Grosse, E. H., Heath, M. T., LeVeque, R. J., Luk, F. T., Nash,S. G., and Trefethen, L. N., *Numerical Analysis Program Library User's Guide: NAPLUG*. SLAC User Note 82, http//www.slac.stanford.edu/spires/... find/hep/www?r=slac-scip-user-note-082, 1979.

[11] Bornemann, F., Laurie, D., Wagon, S., and Waldvogel, J., The SIAM 100-digit challenge, a study in high-accuracy numerical computing. http://www-m3.ma.tum.de/m3old/bornemann/challengebook/index.html.

[12] Bornemann, F., Laurie, D., Wagon, S., and Waldvogel, J., *The SIAM 100-Digit Challenge, A Study in High-Accuracy Numerical Computing*. SIAM, Philadelphia, PA, 2004.

[13] Buckheit, J. B., and Donoho, D. L., WaveLab and reproducible research. http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf, 1995.

[14] Calhoun, D. A., Colella, P., and LeVeque, R. J., CHOMBO-CLAW software. http://www.amath.washington.edu/~calhoun/demos/ChomboClaw.

[15] Calhoun, D. A., Helzel, C., and LeVeque, R. J., Logically Rectangular Grids and Finite Volume Methods for PDEs in Circular and Spherical Domains. In preparation; http://www.amath. washington.edu/~rjl/pubs/circles, 2005.

[16] Colella, P., et al., CHOMBO software. http://seesar.lbl.gov/anag/chombo/, 2005.

[17] Crowder, H., Dembo, R. S., and Mulvey, J. M., On reporting computational experiments with mathematical software. *ACM Trans. Math. Software* **5** (1979), 193–203.

[18] Deiterding, R., AMROC software. http://amroc.sourceforge.net/, 2005.

[19] Deiterding, R., , Construction and application of an amr algorithm for distributed memory computers. In *Adaptive mesh refinement - theory and applications* (ed. by T. Plewa), Lecture Notes in Comput. Sci. Engrg. 41, Springer-Verlag, Berlin 2005, 361–372.

[20] Denlinger, R. P., and Iverson, R. M., Granular avalanches across irregular three-dimensional terrain: 1. Theory and computation. *J. Geophys. Res.* **109** (2004), F01014.

[21] Denlinger, R. P., and Iverson, R. M., Granular avalanches across irregular three-dimensional terrain: 2. Experimental tests. *J. Geophys. Res.* **109** (2004), F01015.

[22] Denlinger, R. P., and O'Connell, D., Two dimensional flow constraints on catastrophic outflow of glacial Lake Missoula over three dimensional terrain. Invited abstract, 3rd International Paleoflood Workshop, Hood River, OR, 2003.

[23] Dongarra, J. J., Bunch, J. R., Moler, C. B., and Stewart, G. W., *LINPACK Users' Guide*. SIAM, Philadelphia, PA, 1979.

[24] Donoho, D. L., and Huo, X., BeamLab and reproducible research. *Int. J. Wavelets Multiresolut. Inf. Process.* **2** (4) (2004), 391–414.

[25] Forsythe, G. E., Galler, B. A., Hartmanis, J., Perlis, A. J., and Traub, J. F., Computer science and mathematics. *ACM SIGCSE Bulletin* **2** (1970), 19–29.

[26] Fritze, J., Extracorporeal shockwave therapy (ESWT) in orthopedic indications: a selective review. *Versicherungsmedizin* **50** (1998), 180–185.

[27] Garbow, B. S., Boyle, J. M. Dongarra, , J. J., and Moler, C. B., *Matrix Eigensystem Routines — EISPACK Guide Extensions*. Lecture Notes in Comput. Sci. 51, Springer-Verlag, Berlin 1977.

[28] Geist, E. L., Titov, V. V., and Synolakis, C. E., Tsunami: wave of change. *Scientific American* **294** (2006), 57–63.

[29] Gentleman, R., and Lang, D. T., Statistical analyses and reproducible research. Bioconductor Project Working Papers. Working Paper 2. http://www.bepress.com/bioconductor/paper2, 2004.

[30] Godunov, S. K., A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb.* **47** (1959), 271–306.

[31] Greenough, J. A., and Rider, W. J., A quantitative comparison of numerical methods for the compressible Euler equations: fifth-order WENO and piecewise-linear Godunov. *J. Comput. Phys.* **196** (2004), 259–281.

[32] Hammer, A. S., Rupp, S., Ensslin, S., Kohn, D., and Seil, R., Extracorporal shock wave therapy in patients with tennis elbow and painful heel. *Archives of Orthopaedic and Trauma Surgery* **120** (2000), 304–307.

[33] Jackson, R. H. F., Boggs, P. T., Nash, S. G., and Powell, S., Guidelines for reporting results of computational experiments. Report of the ad hoc committee. *Math. Programming* **49** (1991), 413–425.

[34] Knuth, D. E., Computer science and its relation to mathematics. Amer. Math. Monthly, 81 (1974), pp. 323–343.

[35] Knuth, D. E., *Algorithmic thinking and mathematical thinking*. *Amer. Math. Monthly* **92** (1985), 170–181.

[36] Langseth, J. O., Wave Propagation Schemes, Operator Splittings, and Front Tracking for Hyperbolic Conservation Laws. PhD thesis, Department of Informatics, University of Oslo, 1996.

[37] Langseth, J. O., and LeVeque, R. J., A wave-propagation method for three-dimensional hyperbolic conservation laws. *J. Comput. Phys.* **165** (2000), 126–166.

[38] Lee, C.-Y., Bard, J., Pinedo, M., and Wilhelm, W. E., Guidelines for reporting computational results in IEE Transactions. *IIE Trans.* **25** (1993), 121–123.

[39] LeVeque, R. J., http://www.amath.washington.edu/~rjl/pubs/icm06.

[40] LeVeque, R. J., CLAWPACK software. http://www.amath.washington.edu/~claw.

[41] LeVeque, R. J., CLAWPACK *User's Guide*. http://www.amath.washington.edu/~claw/doc.html.

[42] LeVeque, R. J., *Numerical Methods for Conservation Laws*. Lectures Math. ETH Zürich, Birkhäuser, Basel 1990.

[43] LeVeque, R. J., *Wave propagation algorithms for multi-dimensional hyperbolic systems*. *J. Comput. Phys.* **131** (1997), 327–353.

[44] LeVeque, R. J., *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts Appl. Math., University Press, Cambridge 2002.

[45] LeVeque, R. J., The dynamics of pressureless dust. *J. Hyperbolic Differential Equations* **1** (2004), 315–327.

[46] Liska, R., and Wendroff, B., Comparison of several difference schemes on 1D and 2D test problems for the Euler equations. *SIAM J. Sci. Comput.* **25** (2003), 996–1017.

[47] Mitran, S., BEARCLAW software. http://www.amath.unc.edu/Faculty/mitran/bearclaw. html.

[48] Moler, C., The origins of MATLAB. MATLAB *News & Notes*, December, 2004. http://www.mathworks.com/company/newsletters/news_notes/clevescorner/dec04.html.

[49] Pelanti, M., Wave Propagation Algorithms for Multicomponent Compressible Flows with Applications to Volcanic Jets. PhD thesis, University of Washington, 2005.

[50] Pelanti, M., and LeVeque, R. J., High-resolution finite volume methods for dusty gas jets and plumes. *SIAM J. Sci. Comput.*, to appear.

[51] Poludnenko, A. Y., Frank, A., and Mitran, S., Clumpy flows in protoplanetary and planetary nebulae. 2003. http://xxx.lanl.gov/abs/astro-ph/0310286.

[52] Poludnenko, A. Y., Frank, A., and Mitran, S., Strings in the Eta Carinae nebula: Hypersonic radiative cosmic bullets. 2003. http://xxx.lanl.gov/abs/astro-ph/0310007.

[53] Roache, P. J., *Verification and Validation in Computational Science and Engineering*. Hermosa Publishers, Albuquerque, NM, 1998.

[54] Rossmanith, J. A., A wave propagation method for hyperbolic systems on the sphere. *J. Comput. Phys.* **213** (2006), 629–658.

[55] Rossmanith, J. A., Bale, D. S., and LeVeque, R. J., A wave propagation algorithm for hyperbolic systems on curved manifolds. *J. Comput. Phys.* **199** (2004), 631–662.

[56] Roy, C. J., Review of code and solution verification procedures for computational simulation. *J. Comput. Phys.* **205** (2005), 131–156.

[57] Schaden, W., Fischer, A., and Sailler, A., Extracorporeal shock wave therapy of nonunion or delayed ossesous union. Clinical Orthopaedics & Related Res. **387** (2001), 90–94.

[58] Schmitt, J., Haake, M., Tosch, A., Hildebrand, R., Deike, B., and Griss, P., Low-energy extracorporeal shock-wave treatment (ESWT) for tendinitis of the supraspinatus. *J. Bone & Joint Surgery* **83-B** (2001), 873–876.

[59] Schwab, M., Karrenbach, M., and Claerbout, J., Making scientific computations reproducible. http://sepwww.stanford.edu/research/redoc/cip.html.

[60] Shu, C.-W., High order ENO and WENO schemes for computational fluid dynamics. In *High-Order Methods for Computational Physics* (ed. by T. J. Barth and H. Deconinck), Lecture Notes in Comput. Sci. Engrg. 9, Springer-Verlag, Berlin 1999, 439–582.

[61] Smith, B. T., Boyle, J. M., Dongarra, J. J., Garbow, B. S., Ikebe, Y., Klema, V. C., and Moler, C. B., *Matrix Eigensystem Routines — EISPACK Guide*. Lecture Notes in Comput. Sci. 6, Springer-Verlag, Berlin 1976.

[62] Trefethen, L. N., *Spectral Methods in Matlab*. SIAM, Philadelphia, PA, 2000.

[63] Trefethen, L. N., *The SIAM 100-digit challenge*. http://www.comlab.ox.ac.uk/nick.trefethen/hundred.html, 2004.

[64] Trefethen, L. N., *Ten digit algorithms*. http://www.comlab.ox.ac.uk/nick.trefethen/ten_digit_algs.htm, 2005.

Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420, U.S.A.

E-mail: rjl@amath.washington.edu

# Reduced basis method for the rapid and reliable solution of partial differential equations

Yvon Maday*

**Abstract.** Numerical approximation of the solution of partial differential equations plays an important role in many areas such as engineering, mechanics, physics, chemistry, biology – for computer-aided design-analysis, computer-aided decision-making or simply better understanding. The fidelity of the simulations with respect to reality is achieved through the combined efforts to derive: (i) better models, (ii) faster numerical algorithm, (iii) more accurate discretization methods and (iv) improved large scale computing resources. In many situations, including optimization and control, the *same* model, depending on a parameter that is changing, has to be simulated over and over, multiplying by a large factor (up to 100 or 1000) the solution procedure cost of one single simulation. The reduced basis method allows to define a surrogate solution procedure, that, thanks to the complementary design of fidelity certificates on outputs, allows to speed up the computations by two to three orders of magnitude while maintaining a sufficient accuracy. We present here the basics of this approach for linear and non linear elliptic and parabolic PDE's.

**Mathematics Subject Classification (2000).** 65D05, 65M60, 65N15, 65N30, 65N35.

**Keywords.** Reduced-basis, a posteriori error estimation, output bounds, offline-online procedures, Galerkin approximation, parametrized partial differential equations.

## 1. Introduction

Let us consider a class of problems depending on a parameter $\mu \in \mathcal{D}$ set in the form: find $u \equiv u(\mu) \in X$ such that $\mathcal{F}(u;\mu) = 0$ (we do not specify much at this point what $\mathcal{D}$ is, it could be a subset of $\mathbb{R}$, or $\mathbb{R}^p$, or even a subset of functions). Such problems arise in many situations such as e.g. optimization, control or parameter-identification problems, response surface or sensibility analysis. In case $\mathcal{F}$ is written through partial differential equations, the problem may be stationary or time dependent but in all these cases, a solution $u(\mu)$ has to be evaluated or computed for many instances of $\mu \in \mathcal{D}$. Even well optimized, the favorite discretization method of yours

*This paper presents a review of results (on the definition, analysis and solution strategies) most of them first presented elsewhere and that have benefitted from the long-standing collaboration with Anthony T. Patera, Einar M. Rønquist, Gabriel Turinici, and more recently Annalisa Buffa, Emil Løvgren, Ngoc Cuong Nguyen, Georges Pau, Christophe Prud'homme. This paper is not intended to be exhaustive and is an invitation to read more complete papers that treat in depth the features presented here. Due to this, no figures nor tables synthesizing numerical results are provided.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

will lead to very heavy computations in order to approximate all these solutions and decision may not be taken appropriately due to too large computer time for reliable simulations.

The approach discussed in this paper will not aim at presenting an alternative to your favorite discretization, more the contrary. The idea is that, in many cases, your discretization will help in constructing a surrogate method that will allow to mimic it or at least to do the spadework on the evaluation of the optimal or control solution. The complexity of the equations resulting from this approach will be very low, enabling very fast solution algorithms. No miracle though, the method is based on a learning strategy concept, and, for a new problem, the preliminary *off-line* preparation is much time consuming. It is only after this learning step that the full speed of the method can be appreciated *on-line*, paying off the cost of the *off-line* preparation step. During the first step, we evaluate accurately, based on your preferred solver, a few solutions to $\mathcal{F}(u;\mu) = 0$; actually, any discretization method is good enough here. In the second step, that is involved on request and *on-line*, the discretization method that has been used earlier is somehow forgotten and a new discretization approach is constructed based on a new ad-hoc basis set (named "reduced basis") built out from the previous computations. In many cases the method proves very efficient and – even though the complete understanding of the reasons why it is working so well are not mastered – an *a posteriori* error theory allows to provide fidelity certificates on outputs computed from the reduced-basis-discretized solution. This method is valid in case the set $\mathcal{S}(\mathcal{D}) = \{u(\mu),\ \mu \in \mathcal{D}\}$ has a simple (hidden) structure, the solution $u(\mu)$ has to be regular enough in $\mu$. We provide some explanations on the rational of the reduced basis approximation in Section 2 and present the method in the elliptic case. In Section 3 we give more rigorous explanation on the rapid convergence of the method on a particular case. This is complemented in Section 4 by an analysis of a posteriori tools that provide fidelity certificate for outputs computed from the reduced basis approximation. Section 5 tells more about the track to follow to be convinced that the method will "work" on the particular problem of yours. The efficient implementation of the reduced basis method needs some care, we present in Section 6 some of the required tools. Finally we end this paper by providing in Section 7 some of the new directions we are currently working on.

## 2. Basics and rational of the reduced basis approach

The reduced basis method consists in approximating the solution $u(\mu)$ of a parameter dependent problem $\mathcal{F}(u;\mu) = 0$ by a linear combination of appropriate, preliminary computed, solutions $u(\mu_i)$ for well chosen parameters $\mu_i, i = 1, \ldots, N$. The rational of this approach, stands in the fact that the set $\mathcal{S}(\mathcal{D}) = \{u(\mu)$ of all solutions when $\mu \in \mathcal{D}\}$ behaves well. In order to apprehend in which sense the good behavior of $\mathcal{S}(\mathcal{D})$ should be understood, it is helpful to introduce the notion of $n$-width following Kolmogorov [8] (see also [14]).

**Definition 2.1.** Let $X$ be a normed linear space, $A$ be a subset of $X$ and $X_n$ be a generic $n$-dimensional subspace of $X$. The deviation of $A$ from $X_n$ is

$$E(A; X_n) = \sup_{x \in A} \inf_{y \in X_n} \|x - y\|_X.$$

The *Kolmogorov n-width* of $A$ in $X$ is given by

$$\begin{aligned} d_n(A, X) &= \inf\{E(A; X_n) : X_n \text{ an } n\text{-dimensional subspace of X}\} \\ &= \inf_{X_n} \sup_{x \in A} \inf_{y \in X_n} \|x - y\|_X. \end{aligned} \tag{1}$$

The $n$-width of $A$ thus measures the extent to which $A$ may be approximated by an $n$-dimensional subspace of $X$. There are many reasons why this $n$-width may go rapidly to zero as $n$ goes to infinity. In our case, where $A = \mathcal{S}(\mathcal{D})$, we can refer to regularity of the solutions $u(\mu)$ with respect to the parameter $\mu$, or even to analyticity. Indeed, an upper bound for the asymptotic rate at which the convergence to zero is achieved is provided by this example from Kolmogorov stating that $d_n(\tilde{B}_2^{(r)}; L^2) = \mathcal{O}(n^{-r})$ where $\tilde{B}_2^{(r)}$ is the unit ball in the Sobolev space of all $2\pi$-periodic real valued, $(r-1)$-times differentiable functions whose $(r-1)$st derivative is absolutely continuous and whose $r$th derivative belongs to $L^2$. Actually, exponential convergence is achieved when analyticity exists in the parameter dependency. The knowledge of the rate of convergence is not sufficient: of theoretical interest is the determination of the (or at least one) optimal finite dimensional space $X_n$ that realizes the infimum in $d_n$, provided it exists. For practical reasons, we want to restrict ourselves to finite dimensional spaces that are spanned by elements of $\mathcal{S}(\mathcal{D})$. This might increase the $n$-width in general Banach space, but of course it does not in Hilbert space as it follows easily from the decomposition of $X$ into $X_A \oplus X_A^{\perp}$, where $X_A$ denotes the vectorial space spanned by $A$. We thus have

$$d_n(A, X_A) = d_n(A, X). \tag{2}$$

We derive from this equality that the quantity

$$\inf \left\{ \sup_{u \in \mathcal{S}(\mathcal{D})} \inf_{y \in X_n} \|x - y\|_X : X_n = \text{Span}\{u(\mu_1), \ldots, u(\mu_n),\ \mu_i \in \mathcal{D}\} \right\} \tag{3}$$

converges to zero (almost at the same speed as $d_n(\mathcal{S}(\mathcal{D}); X)$ provided very little regularity exists in the parameter dependency of the solution $u(\mu)$, and an exponential convergence is achieved in many cases since analyticity in the parameter is quite frequent.

This is at the basics of the reduced basis method. Indeed we are led to choose properly a sequence of parameters $\mu_1, \ldots, \mu_n, \ldots \in \mathcal{D}$, then define the vectorial space $X_N = \text{Span}\{u(\mu_1), \ldots, u(\mu_N)\}$ and look for an approximation of $u(\mu)$ in $X_N$.

Let us consider for example an elliptic problem: Find $u(\mu) \in X$ such that

$$a(u(\mu), v; \mu) = f(v) \quad \text{for all } v \in X. \tag{4}$$

Here $X$ is some Hilbert space, and $a$ is a continuous and elliptic, bilinear form in its two first arguments, regular in the parameter dependance and $f$ is some given continuous linear form. We assume for the sake of simplicity that the ellipticity is uniform with respect to $\mu \in \mathcal{D}$: there exists $\alpha > 0$ such that

$$a(u, u; \mu) \geq \alpha \|u\|_X^2 \quad \text{for all } \mu \in \mathcal{D},\ u \in X,$$

and that the continuity of $a$ is uniform with respect to $\mu \in \mathcal{D}$ as well: there exists $\gamma > 0$ such that

$$|a(u, v; \mu)| \leq \gamma \|u\|_X \|v\|_X \quad \text{for all } \mu \in \mathcal{D},\ u, v \in X.$$

It is classical to state that, under the previous hypothesis, problem (4) has a unique solution for any $\mu \in \mathcal{D}$. The Galerkin method is a standard way to approximate the solution to (4) provided that a finite dimensional subspace $X_N$ on $X$ is given. It consists in: Find $u_N(\mu) \in X_N$ such that

$$a(u_N(\mu), v_N; \mu) = f(v_N) \quad \text{for all } v_N \in X_N, \tag{5}$$

which similarly has a unique solution $u_N(\mu)$. Cea's lemma then states that

$$\|u(\mu) - u_N(\mu)\|_X \leq \left(1 + \frac{\gamma}{\alpha}\right) \inf_{v_N \in X_N} \|u(\mu) - v_N\|_X. \tag{6}$$

The best choice for the basis element $u(\mu_1), \ldots, u(\mu_N)$ of $X_N$ would be those that realize the infimum in (3), i.e. the ones that realize the maximum of the volume $V_N(u(\mu_1), \ldots, u(\mu_N))$ of the parallelepiped determined by the vectors $u(\mu_1), \ldots, u(\mu_N)$. Unfortunately, this is not a constructive method and we generally refer to a greedy algorithm such as the following one:

$$\begin{aligned} \mu_1 &= \arg \sup_{\mu \in \mathcal{D}} \|u(\mu)\|_X, \\ \mu_{i+1} &= \arg \sup_{\mu \in D} \|u(\mu) - P_i u(\mu)\|_X, \end{aligned} \tag{7}$$

where $P_i$ is the orthogonal projection onto $X_i = \text{Span}\{u(\mu_1), \ldots, u(\mu_i)\}$ or a variant of it that is explained at the end of Section 4. The convergence proof for the related algorithm is somehow more complex and presented in a quite general settings in [1].

## 3. An example of *a priori* analysis

The previous notion of $n$-width is quite convenient because it is rather general, in spirit, and provides a tool to reflect the rapid convergence of the reduced basis method but it is not much constructive nor qualitatively informative. We are thus going to consider

a particular example where the parametrized "bilinear" form $a\colon X \times X \times \mathcal{D} \to \mathbb{R}$ is defined as follows

$$a(w, v; \mu) \equiv a_0(w, v) + \mu a_1(w, v); \tag{8}$$

here the bilinear forms $a_0\colon X \times X \to \mathbb{R}$ and $a_1\colon X \times X \to \mathbb{R}$ are continuous, symmetric and positive semi-definite, $\mathcal{D} \equiv [0, \mu_{\max}]$, and we assume that $a_0$ is coercive. It follows from our assumptions that there exists a real positive constant $\gamma_1$ such that

$$0 \leq \frac{a_1(v, v)}{a_0(v, v)} \leq \gamma_1 \quad \text{for all } v \in X. \tag{9}$$

For the hypotheses stated above, it is readily demonstrated that the problem (4) satisfies uniformly the Lax–Milgram hypothesis.

Many situations may be modeled by this rather simple problem statement (4), (8). It can be the conduction in thin plates and $\mu$ represents the convective heat transfer coefficient, it can also be a variable-property heat transfer, then $1 + \mu$ is the ratio of thermal conductivities in domains.

The analysis that we did in [12] involves the eigenproblem: Find $(\varphi \in X, \lambda \in \mathbb{R})$, satisfying $a_1(\varphi, v) = \lambda\, a_0(\varphi, v)$ for all $v \in X$. Indeed the solution $u(\mu)$ to problem (4) can be expressed as

$$u(\cdot, \mu) = \int \frac{f(\varphi)\, \varphi(\cdot; \lambda)}{1 + \mu\lambda} d\lambda. \tag{10}$$

The dependency in $\mu$ is thus explicitly expressed and we can propose to approximate $u(\mu)$ by a linear combination of well chosen $u(\mu_i)$. This can be done through interpolation at the $\mu_i$ by polynomials. It is interesting to notice at this point that we have a large choice in the variable in which the polynomial can be expressed. Indeed since we are interested through this interpolation process to evaluate the best fit, a polynomial in $\mu$ may not be the best choice but rather a polynomial in $\frac{1}{\mu}$, $e^{\mu}$ or else; in [12] we have considered a polynomial approximation in the variable $\tau = \ln(\mu + \delta^{-1})$, where $\delta$ is some positive real number. The analysis then involves the interpolation operator at equidistant points (in the variable $\tau$) for which we were able to get an upper bound used, in turn, to qualify the best fit result

**Lemma 3.1.** *There exists a constant $C > 0$ and a positive integer $N_{\text{crit}}$ such that for $N \geq N_{\text{crit}}$*

$$\inf_{w_N \in X_N} \|u(\mu) - w_N\|_X \leq C \exp\left\{\frac{-N}{N_{\text{crit}}}\right\} \quad \textit{for all } \mu \in \mathcal{D},$$

*where $N_{\text{crit}} \equiv c^* e\, \ln(\gamma\, \mu_{\max} + 1)$.*

This analysis of [12] leads to at least three remarks:

**Remark 3.2.** a) The analysis of the best fit done here suggests to use sample points $\mu_i$ that are equidistant when transformed in the $\tau$ variable. We performed some numerical

tests to check whether this sampling gives indeed better results than more conventional ones (of course you should avoid equidistant in the original $\mu$ variable, but we tried e.g. Chebyshev points) and this was actually the case. Unfortunately, in more general situations and especially in higher parameter dimensions, we have no clue of a direct constructive best sampling method.

b) For a given sampling $\mu_i$, one can propose an interpolation procedure to approximate $u(\mu)$ which is more simple than referring to a Galerkin approach. Indeed, an approximation

$$u(\mu) \simeq \sum_{i=1}^{N} \alpha_i(\mu) u(\mu_i),$$

can be proposed by using coefficients that are the Lagrange interpolation basis in the chosen variable (above it was $\tau = \ln(\mu + \delta^{-1})$, i.e. the mapping $\tau \mapsto \alpha_i(\mu(\tau))$ is a polynomial of degree $\leq N$ and $\alpha_i(\mu_j) = \delta_{ij}$). The problem is that the expression of $\alpha_i(\mu)$ depends on the best choice of variable which is unknown and within a set that is quite infinite providing a range of results that are quite different. Since for a given general problem we have no clue of the best interpolation system, the Galerkin approach makes sense, indeed.

c) In opposition, the Galerkin approach does not require any preliminary analysis on guessing the way the solution depends upon the parameter. Its superiority over interpolation process comes from the fact stated in Cea's lemma that the approximation that is obtained, up to some multiplicative constant, gives the optimal best fit, even if we do not know the rate at which the convergence is going.

Finally, as is often the case, we should indicate that the a priori analysis helps to have confidence in developing the method but, at the end of a given computation, a computable a posteriori estimator should be designed in order to qualify the approximation. This is even more true with such new surrogate approximation in order to replace the expertise a user may have in his preferred method, e.g. his intuition on the choice of the discretization parameter to get acceptable discrete solutions. This is the purpose of the next section.

## 4. An example of *a posteriori* analysis

Most of the time, the complete knowledge of the solution of the problem (4) is not required. What is required, is outputs computed from the solution $s = s(u)$, where $s$ is some continuous functional defined over $X$. In order to have a hand over this output, the reduced basis method consists first in computing $u_N \in X_N$ solution of the Galerkin approximation (5), then propose $s_N = s(u_N)$ as an approximation of $s$. Assuming Lipschitz condition (ex. linear case) over $s$, it follows that

$$|s - s_N| \leq c\|u - u_N\|_X. \tag{11}$$

Thus any information over the error in the energy norm will allow to get verification (provided you are able to evaluate $c$). Actually it is well known that the convergence of $s_N$ towards $s$ most often goes faster. This is standard but we go back over it since this will prove useful in the sequel. Let us assume we are in the linear output case where $s \equiv \ell$ is a linear continuous mapping over $X$. It is then standard to introduce the *adjoint state*, solution of the following problem: find $\psi \in X$ such that

$$a(v, \psi; \mu) = -\ell(v) \quad \text{for all } v \in X. \tag{12}$$

The error in the output is then (remember that, for any $\phi_N \in X_N$, $a(u, \phi_N; \mu) = a(u_N, \phi_N; \mu) = (f, \phi_N)$)

$$\begin{aligned} s_N - s &= \ell(u_N) - \ell(u) \\ &= a(u, \psi; \mu) - a(u_N, \psi; \mu) \\ &= a(u, \psi - \phi_N; \mu) - a(u_N, \psi - \phi_N; \mu) \quad (\text{for all } \phi_N \in X_N) \\ &= a(u - u_N, \psi - \phi_N; \mu) \quad (\text{for all } \phi_N \in X_N) \\ &\leq c\|u - u_N\|_X \|\psi - \phi_N\|_X \quad (\text{for all } \phi_N \in X_N), \end{aligned} \tag{13}$$

so that the best fit of $\psi$ in $X_N$ can be chosen in order to improve the first error bound (11) that was proposed for $|s - s_N|$.

For instance if $\psi_N$ is the solution of the Galerkin approximation to $\psi$ in $X_N$, we get

$$|s - s_N| \leq c\|u - u_N\|_X \|\psi - \psi_N\|_X. \tag{14}$$

Actually, the approximation of $\psi$ in $X_N$ may not be very accurate since $X_N$ is well suited for approximating the elements $u(\mu)$ and – except in the case where $\ell = f$ named the compliant case – a separate reduced space $\tilde{X}_N$ should be built which provides an associated approximation $\tilde{\psi}_N$. Then an improved approximation for $\ell(u)$ is given by $\ell_{\text{imp}} = \ell(u_N) - a(u_N, \tilde{\psi}_N) + f(\tilde{\psi}_N)$ since (14) holds with $\|\psi - \tilde{\psi}_N\|_X$ for which a better convergence rate is generally observed.

Even improved, this result is still *a priori* business and it does not allow to qualify the approximation for a given computation. In order to get *a posteriori* information, between $\ell(u)$ and $\ell(u_N)$ (or $\ell_{\text{imp}}$), we have to get a hand on the residuals in the approximations of the primal and dual problems. We introduce for any $v \in X$,

$$\mathcal{R}^{\text{pr}}(v; \mu) = a(u_N, v; \mu) - \langle f, v \rangle, \quad \mathcal{R}^{\text{du}}(v; \mu) = -a(v, \tilde{\psi}_N; \mu) - \ell(v). \tag{15}$$

We then compute the reconstructed errors associated with the previous residuals. These are the solutions of the following problems:

$$2\alpha \int \nabla \hat{e}^{\,\text{pr(du)}} \nabla v = \mathcal{R}^{\text{pr(du)}}(v; \mu) \quad \text{for all } v. \tag{16}$$

We then have

**Theorem 4.1.** *Let* $s^- = \ell_{\rm imp} - \alpha \int [\nabla(\hat{e}^{\,\rm pr} + \hat{e}^{\,\rm du})]^2$ *then* $s^- \le s$. *In addition, there exists two constants* $0 < c \le C$ *such that*

$$c|s - s_N| \le s - s^- \le C|s - s_N|.$$

*Proof.* Let us denote by $e_N$ the difference between the exact solution and the approximated one $e_N = u - u_N$. From (16), we observe that

$$2\alpha \int \nabla \hat{e}^{\rm pr} \nabla e_N = -a(e_N, e_N; \mu)$$

and

$$2\alpha \int \nabla \hat{e}^{\rm du} \nabla e_N = -a(e_N, \tilde{\psi}_N; \mu) - \ell(e_N) = f(\tilde{\psi}_N) - a(u_N, \tilde{\psi}_N) - \ell(e_N).$$

Taking this into account allows to write

$$\begin{aligned} \ell_{\rm imp} - \alpha \int \nabla(\hat{e}^{\rm pr} + \hat{e}^{\rm du})^2 &= \ell(u_N) - a(u_N, \tilde{\psi}_N) + f(\tilde{\psi}_N) - \alpha \int \nabla(\hat{e}^{\rm pr} + \hat{e}^{\rm du})^2 \\ &= \ell(u) - \alpha \int \nabla(e_N + \hat{e}^{\rm pr} + \hat{e}^{\rm du})^2 - a(e_N, e_N; \mu) + \alpha \int [\nabla e_N]^2, \end{aligned} \tag{17}$$

and the proof follows from the uniform ellipticity of $a(\,.\,,.\,;\mu)$. □

Despite the fact that we have avoided to speak about any discretization so far, Theorem 4.1 is already informative in the sense that in order to obtain $s^-$, the problem (16) to be solved, is parameter independent and simpler than the original one, provided that we have a good evaluation of the ellipticity constant. In Section 6 we shall explain how to transform these constructions in a method that can be implemented. Before this we should explain how the previous estimator may help in designing a good choice for the elements of the reduced basis, providing a third alternative to the greedy algorithm presented in (7). Currently indeed, we have two alternative, either a random approach (that generally works not so badly) or select out of a large number of pre-computed solution $\{u_i\}_i$, the best sample from a SVD approach by reducing the matrix of scalar products $(u_i, u_j)$. The former lacks of fiability, the latter is a quite expensive approach and is mostly considered in a pre analysis framework as is explained in the next section. In order to reduce the cost of the off-line stage we can propose a greedy algorithm that combines the reduced approximation and the error evaluation:

- Take a first parameter (randomly).
- Use a (1-dimensional) reduced basis approach over a set of parameter values (chosen randomly) and select, as a second parameter, the one for which the associated predicted error $s^+ - s^-$ is the largest.

This gives now a 2-dimensional reduced basis method.

- Use this (2-dimensional) reduced basis approach over the same set of parameters and select, as a third parameter, the one for which the associated error is the largest.

This gives a 3-dimensional reduced basis method ...

- and proceed ...

Note that we then only compute accurately the solutions corresponding to the parameters that are selected this way.

The a posteriori approach that has been presented above relies on the uniform ellipticity of the bilinear form and the knowledge of the ellipticity constant. For more general problems, where only, nonuniform inf-sup conditions are valid (e.g. the noncoercive Helmholtz acoustics problem which becomes singular as we approach resonance) smarter definitions should be considered. We refer to [18] for improved methods in this direction.

## 5. Some pragmatic considerations

Now that some basics on the reduced basis method have been presented, it is interesting to understand if the problem you have in mind is actually eligible to this type of approximation. We are thus going to propose some pragmatic arguments that may help in the preliminary verification. First of all, let us note that we have illustrated the discretization on linear elliptic problems, of course this is just for the sake of simplicity, non linear problem [11], [19], [20] so as time dependent problems [7], [17] can be solved by these methods. Second, many things can be considered as a valid parameter: this can be the size of some simple geometric domain on which the solution is searched [16] , but it can be the shape itself [13] (the parameter in the former case is a multireal entity while in the latter it is a functional), the parameter can also be the time [7], [17] , or the position of some given singularities [2].

For all these choices, a fair regularity in the parameter is expected and wished so that the $n$-width goes fast to zero. An important remark has to be done here in order the size of the reduced basis be the smallest possible. Indeed, it may be smart to preprocess the precomputed solutions in order they look more similar. An example is given in [2] where quantum problem are considered; the solutions to these problems present some singularities at the position of the nuclei. If the position of the nuclei is the parameter we consider, it is useful to transform each solution in a reference configuration where the singularities/nuclei are at a unique place; the solutions are then much more comparable. Another example is given by the solution of the incompressible Stokes and Navier–Stokes problem where the shape of the computational domain is the parameter; in order to be able to compare them properly, they have to be mapped on a unique (reference) domain. This is generally done through a simple change of variable. In case of the velocity, it is a vector field that is divergence

free and a "standard" change of variable will (generally) not preserve this feature. The Piola transform (that actually corresponds to the classical change of variable over the potential function) allows to have the velocity fields transformed over the reference domain while preserving the divergence free condition as is demonstrated in [9]. These preprocessing steps allow to diminish the $n$-width of $\mathcal{S}(\mathcal{D})$ and it pays to be smart!

In order to sustain the intuition on the potential of the reduced basis concept, a classical way is to use a SVD approach. Let us assume that we have a bunch of solutions $u_i = u(\mu_i)$, snapshots of the space $\mathcal{S}(\mathcal{D})$ of solutions to our problem. Out of these, the correlation matrix $(u_i, u_j)$ which is symmetric can be reduced to its eigen-form, with positive eigenvectors that, ranked in decreasing order, go to zero. The high speed of convergence towards zero of the eigenvalues ranked in decreasing order will sustain the intuition that the reduced basis method will work. Indeed, the $n$-width is directly related to the size of the eigenvalues larger than the $n+1$th. The idea is that if the number of eigenvectors associated with the largest eigenvalues is small, then the method is viable. In order to sustain this, you can also consider, momentarily, the space $X_N$ spanned by the eigenvectors associated with the $N$ largest eigenvalues and analyze the norm of the difference between the snapshots in $\mathcal{S}(\mathcal{D})$ and their best fit in $X_N$. Note that we do not claim that this is a cheap constructive method: this procedure consists in a pre-analysis of the potential of the reduced basis method to approximate the problem you consider. If the norm of the error goes to zero sufficiently fast, you know that a Galerkin approach will provide the same order of convergence and the method is worth trying. We insist on the fact that this pre-analysis is not mandatory, it is only to help in understanding what you should expect, "at best" from the reduced basis approximation. In particular the greedy approach presented in Section 4 has to be preferred to the SVD approach that we discussed above for the determination of the elements that are to be incorporated in the reduced basis space, if you do not want to spend too much time during the off-line stage. Note also that the greedy approach provides solutions, that, when their number becomes large, become more and more linearly dependent (actually this is one of the aspects of the low $n$-width) and thus, for stability purposes it is important, through a Gram–Schmidt process, to extract, from these solutions, orthonormal elements that will be the actual elements of the reduced basis: these will be named $(\zeta_i)_i$. This does not change the potential approximation properties of the reduced basis but improves, to a large extent, the stability of the implementation.

Finally, the preselection may be quite generous in the sense that you may be interested to select more than $N$ basis functions, $N$ being an evaluation of the dimension of the reduced basis for most problems. The reason for this comes from the conclusion of the a posteriori analysis that may tell you to increase the size of the reduced basis, suggesting you to work with $N+2$ (say) instead of $N$ basis functions. This again is a feature of exponentially rapid convergence that lead to a large difference between the accuracy provided by $X_N$ and $X_{N+2}$ (say). It is time now to give some details on the implementation of the method.

## 6. Implementation issues

We start by emphasizing that any reduced basis method necessarily involves the implementation of a more "classical" approximation method. Indeed – except for very particular and uninteresting problems – the knowledge of the solutions, that we named $u_i$, is impossible without referring to a discretization method (e.g. of finite element, spectral type, etc.). This is also the case for the $\zeta$ that are coming out from some shaping of the basis, e.g. Gram–Schmidt, as explained earlier. This is the reason why reduced basis methods should not be considered as competitor to standard approximation methods but only as surrogate approaches.

This implies, though, some difficulties since the elements of the *reduced* basis are only known through a preliminary *computation* basis, which, if we want the solution $u_i$ to be well approximated, has to be *very* large. Knowing this, the rule of the game for the efficient implementation of any reduced basis method is to strictly prohibit any *online* reference to the extended basis. We allow *offline* precomputations of the solutions (that involves the extended basis) and some *offline* cross contribution of these solutions (based on their expression with respect to the extended basis) but this is forbidden *online*. Following [16], we explain in the next subsection how this can be done.

**6.1. Black box approach.** The solution procedure involves the evaluation of the elements of the stiffness matrix $a(\zeta_i, \zeta_j; \mu), 1 \leq i, j \leq N$ that depends on the current parameter $\mu$. This computation involves some derivatives and the evaluation of integrals, that have to be performed and this may be *very* lengthy. It should be stated here that the implementation of the reduced type method has to be much faster than the solution procedure that was used to compute the reduced basis, much means many order of magnitude. The $\mathcal{O}(\dim X_N)^2$ entrees of the stiffness matrix have thus to be evaluated through some smart way.

Let us begin by the easy case that is named *affine parametric dependance* where the entries $a(\zeta_i, \zeta_j; \mu)$ appear to read

$$a(\zeta_i, \zeta_j; \mu) = \sum_p g_p(\mu) a_p(\zeta_n, \zeta_m), \tag{18}$$

where the bilinear forms $a_p$ are parameter independent. Many simple problems where the parameter are local constitutive coefficients or local zooming isotropic or non isotropic factors, enter in this framework. The expensive computation of the $a_{p,n,m} = a_p(\zeta_n, \zeta_m)$ can be done offline, once the reduced basis is constructed; these $a_{p,n,m}$ are stored and, for each new problem, the evaluation of the stiffness matrix is done, online, in $P \times N^2$ operations, and solved in $\mathcal{O}(\dim X_N^3)$ operations. These figures are coherent with the rapid evaluation of the reduced basis method.

**6.2. A posteriori implementation.** Under the same affine dependance hypothesis on $a$, it is easy to explain how the a posteriori analysis can be implemented, re-

sulting in a fast on-line solution procedure, provided some off-line computations are made. First of all the computation of $\tilde{\psi}_N$ can be implemented in the space $\tilde{X}_N = \text{Span}\{\xi_1, \ldots, \xi_N)\}$ exactly as above for the computation of $u_N$. Taking into account (18), together with the expressions obtained from the inversion of problem (5) and (12): $u_N = \sum_{i=1}^N \alpha_i \zeta_i$ and $\tilde{\psi}_N = \sum_{i=1}^N \tilde{\alpha}_i \xi_i$, we can write

$$\mathcal{R}^{\text{pr}}(v, \mu) = \sum_p \sum_i g_p(\mu)\alpha_i a_p(\zeta_i, v) - (f, v),$$

and

$$\mathcal{R}^{\text{du}}(v, \mu) = -\sum_p \sum_j g_p(\mu)\tilde{\alpha}_j a_p(v, \xi_j) - \ell(v),$$

hence by solving *numerically*, off-line, each of the problems

$$2\alpha \int \nabla e^{\text{pr},p,i} \nabla v = a_p(\zeta_i, v), \tag{19}$$

$$2\alpha \int \nabla e^{\text{pr},0} \nabla v = (f, v), \tag{20}$$

$$2\alpha \int \nabla e^{\text{du},p,j} \nabla v = a_p(v, \xi_j), \tag{21}$$

$$2\alpha \int \nabla e^{\text{du},0} \nabla v = \ell(v) \tag{22}$$

allows to write the numerical solutions of (16) as a linear combination of the elements previously computed (e.g. $\hat{e}^{\,\text{pr}} = \sum_p \sum_i g_p(\mu)\alpha_i e^{\text{pr},p,i} - e^{\text{pr},0}$) in $\mathcal{O}(PN)$ operations.

**6.3. Magic points.** The hypothesis of *affine parametric dependency* is rather restrictive, and has to be generalized. In case of quadratic or cubic dependency, this is quite straightforward but even for linear problems such as the Laplace problem, when e.g. geometry is the parameter, this is rarely the case and another approach has to be designed. In order to get a better understanding of the method, let us first indicate that, when the geometry is the parameter, the solutions have to be mapped over a reference domain $\hat{\Omega}$. Let us assume that we want to compute $d(\zeta_i, \zeta_j; \Omega)$ where

$$d(u, v; \Omega) = \int_\Omega uv \, dA = \int_{\hat{\Omega}} uv J_\Phi \, d\hat{A},$$

where $J_\Phi$ is a Jacobian of the transformation that maps $\hat{\Omega}$ onto $\Omega$. There is no reason in the general case that $J_\Phi$ will be affine so that the previous approach will not work. It is nevertheless likely that there exists a sequence of well chosen transformations $\Phi_1^*, \ldots, \Phi_M^*, \ldots$, such that $J_\Phi$ may be well approximated by an expansion $J_\Phi \simeq$

$\sum_{j=1}^{M} \beta_j J_{\Phi_j^*}$. An approximation of $d(\zeta_i, \zeta_j; \Omega)$ will then be given by

$$d(\zeta_i, \zeta_j; \Omega) \simeq \sum_{j=1}^{M} \beta_j \int_{\hat{\Omega}} \hat{\zeta}_i \hat{\zeta}_j J_{\Phi_j^*} \, d\hat{A}, \tag{23}$$

and again, the contributions $\int_{\hat{\Omega}} \hat{\zeta}_i \hat{\zeta}_j J_{\Phi_j^*} d\hat{A}$ will be pre-computed offline. We do not elaborate here on how the $\Phi_j^*$ are selected, and refer to [9], what we want to address is the evaluation of the coefficients $\beta_j = \beta_j(\Omega)$ in the approximation of $J_\Phi$ above. The idea is to use an interpolation procedure as is explained in [6]. Let $\boldsymbol{x}_1$ be the point where $|J_{\Phi_1^*}|$ achieves its maximum value. Assuming then that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ have been defined, and are such that the $n \times n$ matrix with entries $J_{\Phi_k^*}(\boldsymbol{x}_\ell)$, $1 \leq k, \ell \leq n$ is invertible, we define $\boldsymbol{x}_{n+1}$ as being the point where $r_{n+1} = |J_{\Phi_{n+1}^*} - \sum_{k=1}^{n} \gamma_k J_{\Phi_k^*}|$ achieves it maximum value, here the scalar $\gamma_k$ are defined so that $r_{n+1}$ vanishes at any $(\boldsymbol{x}_\ell)$ for $\ell = 1, \ldots, n$. The definition of the points $\boldsymbol{x}_k$ is possible as long the $\Phi_j$ are chosen such that the $J_{\Phi_k^*}$ are linearly independent (see [6]). The $\beta_j$ are then evaluated also through the interpolation process

$$J_\Phi(\boldsymbol{x}_\ell) = \sum_{k=1}^{M} \beta_k J_{\Phi_k^*}(\boldsymbol{x}_\ell) \quad \text{for all } 1 \leq \ell \leq M. \tag{24}$$

We have not much theory confirming the very good results that we obtain (which makes us call these interpolation point "magic"). An indicator that allows to be quite confident in the interpolation process is the fact that the Lebesgue constant attached to the previously built points is, in all example we have encountered, is rather limited.

Note that the same interpolation approach allows to compute the reconstructed errors with a compatible complexity as in the previous subsection.

The same magic point method has to be used also for the implementation of the reduced basis method for nonlinear problems. Actually, denoting by $z_i = \mathrm{NL}(u_i)$ the nonlinear expression involved in the problem, provided that the set $Z_M = \mathrm{Span}\{z_i, 1 \leq i \leq M\}$ has a small width, the interpolation process presented above allows both to determine a good interpolation set and a good associated interpolation nodes, we refer to [6] for more details on the implementation and to numerical results.

## 7. Some extensions

**7.1. Eigenvalue problems.** We end this paper by noticing that the reduced basis method can actually be found, at least in spirit, in many other approximations. There are indeed many numerical approaches that, in order to tackle a complex problem, use the knowledge of the solution of similar but simpler problems to facilitate the approximation. In this direction, the modal synthesis method provides a method to solve approximately eigenvalue problems on large structures based on the knowledge

of the eigenvalues and eigenfunctions of the same problem on substructures. We refer e.g. to [4], [5] for more details on a high order implementation of these approaches.

Similarly, again, one of the approaches for the resolution of Hartree–Fock problem in quantum chemistry is the L.C.A.O. method that consists in approximating the wave function of a molecule by linear combination of atomic orbitals that are nothing but solutions to the same problem on an atom, instead of a molecule. The atomic orbitals are indeed the approximations of Hydrogenoid functions (the contracted Gaussians have to be seen this way) that are the solutions of the electronic problem of one electron around a nucleus. This similarity is the guideline for the extension that is proposed in [3], [2].

At this level, it is also interesting to note that the reduced basis method, for an eigenvalue problem as the one encountered in the two previous cases, may be very appropriate since it can be proven that, letting $u_i$ denote the set of all first $P$ eigenvectors of an eigenvalue problem depending on a parameter $\mu$, $u_i \equiv (e^1(\mu_i), \ldots, e^p(\mu_i))$, then the approximation of this complete set of eigenvectors can be done with the same linear combination. More precisely it is possible to get an accurate approximation method based on

$$u(\mu) \simeq \sum_{i=1}^{P} \alpha_i u_i, \quad e^j(\mu) \simeq \sum_{i=1}^{P} \alpha_i e^j(\mu_i) \text{ for all } j$$

instead of

$$e^j(\mu) \simeq \sum_{i=1}^{P} \alpha_i^j e^j(\mu_i).$$

Again we refer to [2] for more details on this.

**7.2. The reduced element method.** In the reduced basis element method introduced in [13], we consider the geometry of the computational domain to be the generic parameter. The domain is decomposed into smaller blocks, all of them can be viewed as the deformation of a few reference shapes. Associated with each reference shape are previously computed solutions (typically computed over different deformations of the reference shapes). The precomputed solutions are mapped from the reference shapes to the different blocks of the decomposed domain, and the solution on each block is found as a linear combination of the mapped precomputed solutions. The solutions on the different blocks are glued together using Lagrange multipliers.

To be more precise, we assume that the domain $\Omega$ where the computation should be performed can be written as the *non-overlapping* union of subdomains $\Omega^k$:

$$\overline{\Omega} = \bigcup_{k=1}^{K} \overline{\Omega}^k, \quad \Omega^k \cap \Omega^\ell = \emptyset \text{ for } k \neq \ell. \tag{25}$$

Next, we assume that each subdomain $\Omega^k$ is the deformation of the "reference" domain $\hat{\Omega}$ through a regular enough, and one to one, mapping. In an off-line stage,

this reference geometry has been "filled up" with reference reduced basis solutions $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N$ to the problem that is under evaluation. Hence, together with this geometric decomposition, a functional decomposition is proposed since every $\Omega^k$; this allows us to define the finite dimensional space

$$Y_N = \left\{ v \in L^2(\Omega),\ v_{|\Omega^k} = \sum_{i=1}^{N} \alpha_i^k \mathcal{F}_k^{-1}[\hat{u}_i] \right\}, \tag{26}$$

which is a set of uncoupled, element by element, discrete functions, where $\mathcal{F}_k$ allows to transform functions defined over $\hat{\Omega}$ into functions defined over $\Omega^k$. This is generally not yet adequate for the approximation of the problem of interest since some glue at the interfaces $\gamma_{k,\ell}$ between two adjacent domains $\overline{\Omega}^k \cap \overline{\Omega}^\ell$ has to be added to the elements of $Y_N$, the glue depending on the type of equations we are interested to solve (it will be relaxed $C^0$-continuity condition for a Laplace operator, or more generally relaxed $C^1$-continuity condition for a fourth-order operator.

At this stage it should be noticed that, modulo an increase of complexity in the notations, there may exist not only one reference domain $\hat{\Omega}$ filled with its reduced basis functions but a few numbers so that the user can have more flexibilities in the design of the final global shape by assembling deformed basic shapes like a plumber would do for a central heating installation.

The reduced basis element method is then defined as a Galerkin approximation over the space $X_N$ being defined from $Y_N$ by imposing these relaxed continuity constraints. We refer to [9], [10] for more details on the implementation for hierarchical fluid flow systems that can be decomposed into a series of pipes and bifurcations.

## References


[1] Buffa, A., Maday, Y., Patera, A. T., Prud'homme, C., Turinici, G. In progress, 2006.

[2] Cancès, E., Le Bris, C., Maday, Y., Nguyen, N. C., Patera, A. T., Pau, G., Turinici, G. In progress, 2006.

[3] Cancès, E., Le Bris, C., Maday, Y., Turinici, G., Towards reduced basis approaches in ab initio electronic structure computations. *J. Sci. Comput.* **17** (1–4) (2002), 461–469.

[4] Charpentier, I., Maday, Y., Patera, A. T., Bounds evaluation for ouputs of eigenvalue problems approximated by the overlapping modal synthesis method. *C. R. Acad. Sci. Paris Sér. I Math.* **329** (1999), 909–914.

[5] Charpentier, I., de Vuyst, F. , Maday, Y., The overlapping component mode synthesis: The shifted eigenmodes strategy and the case of self-adjoint operators with discontinuous coefficients. In *Proceedings of the Ninth International Conference on Domain Decomposition Methods* (ed. by P. E. Bjørstad, S. Magne, S. Espedal and D. E. Keyes), 1998, 583–596.

[6] Grepl, M. A., Maday, Y., Nguyen, N. C., Patera, A. T., Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *Math. Model. Numer. Anal.*, submitted.

[7] Grepl, M. A., Patera, A. T., A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *Math. Model. Numer. Anal.* **39** (1) (2005), 157–181.

[8] Kolmogoroff, A., Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. *Ann. of Math.* **37** (1963), 107–110.

[9] Løvgren, A. E., Maday, Y., Rønquist, E. M., A reduced basis element method for the steady Stokes problem. *Math. Model. Numer. Anal.*, to appear.

[10] Løvgren, A. E. , Maday, Y. , Rønquist, E. M., The reduced basis element method for fluid flows. *Adv. Math. Fluid Mech.*, 2006, to appear.

[11] Machiels, L. , Maday, M., Oliveira, I. B., Patera, A. T., Rovas, D. V., Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris Sér. I Math.* **331** (2000), 153–158.

[12] Maday, Y., Patera, A. T., Turinici, G., A Priori Convergence Theory for Reduced-Basis Approximations of Single-Parameter Elliptic Partial Differential Equations. *J. Sci. Comput.* **17** (1–4) (2002), 437–446.

[13] Maday, Y., and Rønquist, E. M., The reduced-basis element method: Application to a thermal fin problem. *SIAM J. Sci. Comput.* **26** (2004), 240–258.

[14] Pinkus, A., *n-Widths in Approximation Theory*. Ergeb. Math. Grenzgeb. (3) 7, Springer-Verlag, Berlin 1985.

[15] Prud'homme, C., Contributions aux simulations temps-réel fiables et certains aspects du calcul scientifique. Mémoire d'Habilitation à Diriger les Recherches, Université Pierre et Marie Curie, Paris, 2005.

[16] Prud'homme, C., Rovas, D. V., Veroy, K., Machiels, L., Maday, Y., Patera, A. T., Turinici, G., Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *J. Fluids Engrg.* **124** (2002), 70–80.

[17] Rovas, D. V., Machiels, L., Maday, Y., Reduced-basis output bound methods for parabolic problems. *IMA J. Numer. Anal.*, Advance Access published on March 6, (2006).

[18] Sen, S., Veroy, K., Huynh, D. B. P., Deparis, S., Nguyen, N. C., Patera, A. T., Natural norm, a posteriori error estimators for reduced basis approximations. J. Comp. Phys., 2006, to appear.

[19] Veroy, K., Prud'homme, C., Patera, A. T., Reduced-basis approximation of the viscous Burgers equation: Rigorous a posteriori error bounds. *C. R. Acad. Sci. Paris Ser. I Math.* **337** (2003), 619–624.

[20] Veroy, K., Patera, A. T., Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: Rigorous reduced-basis a posteriori error bounds. *Int. J. Numer. Meth. Fluids* **47** (2005), 773–788.

Université Pierre et Marie Curie-Paris6, UMR 7598 Laboratoire Jacques-Louis Lions,
B.C. 187, 75005 Paris, France
and
Division of Applied Mathematics, Brown University, Providence, U.S.A.
E-mail: maday@ann.jussieu.fr

# Finite element algorithms for transport-diffusion problems: stability, adaptivity, tractability

Endre Süli

**Abstract.** Partial differential equations with nonnegative characteristic form arise in numerous mathematical models of physical phenomena: stochastic analysis, in particular, is a fertile source of equations of this kind. We survey recent developments concerning the finite element approximation of these equations, focusing on three relevant aspects: (a) stability and stabilisation; (b) *hp*-adaptive algorithms driven by residual-based *a posteriori* error bounds, capable of automatic variation of the granularity $h$ of the finite element partition and of the local polynomial degree $p$; (c) complexity-reduction for high-dimensional transport-diffusion problems by stabilised sparse finite element methods.

**Mathematics Subject Classification (2000).** Primary 65N30; Secondary 65N12, 65N15.

**Keywords.** Transport-dominated diffusion problems, Fokker–Planck equations, finite element methods, stability, a-posteriori error analysis, adaptivity, sparse finite elements.

## 1. Introduction

Let $\Omega$ be a bounded and simply-connected open set in $\mathbb{R}^d$, $d \geq 2$, with Lipschitz continuous boundary $\partial\Omega$. On $\Omega$, we consider the partial differential equation

$$\mathcal{L}u := -\nabla \cdot (a\nabla u) + \nabla \cdot (bu) + cu = f, \tag{1}$$

where $f \in \mathrm{L}_2(\Omega)$ and $c \in \mathrm{L}_\infty(\Omega)$ are real-valued functions, $b = \{b_i\}_{i=1}^d$ is a vector function whose entries $b_i$ are Lipschitz continuous real-valued functions on $\overline{\Omega}$. We shall, further, assume that $a = \{a_{ij}\}_{i,j=1}^d$ is a *symmetric* matrix whose entries $a_{ij}$ are bounded, Lipschitz continuous real-valued functions defined on $\overline{\Omega}$, and that the matrix $a$ is positive semidefinite, almost everywhere on $\overline{\Omega}$, i.e.,

$$\alpha(\xi) := \xi^\top a(x)\xi \geq 0 \quad \text{for all } \xi \in \mathbb{R}^d \text{ and a.e. } x \in \overline{\Omega}. \tag{2}$$

Under hypothesis (2), the equation (1) is referred to as a *partial differential equation with nonnegative characteristic form*. Equations of this kind frequently arise as mathematical models in physics and chemistry [40] (e.g. in the kinetic theory of polymers [7], [44], [49] and coagulation-fragmentation problems [43]). They also appear in molecular biology [21], population genetics (e.g. in mathematical models of random

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

genetic drift) and in mathematical finance. Important special cases of these equations include the following: (a) when the diffusion matrix $a = a^\top$ is positive definite, (1) is an elliptic partial differential equation; (b) when $a \equiv 0$ and the transport direction $b \neq 0$, the partial differential equation (1) is a first-order hyperbolic equation; (c) when $b = (0, \ldots, 0, 1)^\top \in \mathbb{R}^d$ and

$$a = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix}$$

where $\alpha$ is a $(d-1) \times (d-1)$ symmetric positive definite matrix, (1) is a parabolic partial differential equation, with time-like direction $b$. The family of partial differential equations with nonnegative characteristic form also includes a range of other linear second-order partial differential equations, such as degenerate elliptic and ultra-parabolic equations. Furthermore, by a result of Hörmander [35] (cf. Theorem 11.1.10 on p. 67), second-order hypoelliptic operators with constant coefficients have nonnegative characteristic form, after possible multiplication by $-1$.

For classical types of partial differential equations, such as those under (a), (b) and (c) above, rich families of reliable, stable and accurate numerical techniques have been developed. Yet, there have only been isolated attempts to date to explore computational aspects of the class of partial differential equations with nonnegative characteristic form as a whole (cf. [30] and [33]). In particular, only a limited amount of research has been devoted to the construction and mathematical analysis of adaptive finite element algorithms for these equations; similarly, there has been very little work on the finite element approximation of high-dimensional partial differential equations with nonnegative characteristic form (cf. [57]).

The aim of this paper is to present a brief survey of some recent results in these directions. In Section 2, we state the weak formulation of a boundary-value problem for equation (1). In Sections 3 and 4, we give an overview of stabilised continuous and discontinuous finite element approximations to these equations; we shall also address the question of residual-based a posteriori error estimation for $hp$-version discontinuous Galerkin approximations of these equations, as well as the construction of sparse stabilised finite element methods for *high-dimensional* partial differential equations with nonnegative characteristic form (Section 5).

## 2. Boundary conditions and weak formulation

For the sake of simplicity of presentation, we shall assume that $\Omega$ is a bounded open polytope in $\mathbb{R}^d$ and we denote by $\Gamma$ the union of its $(d-1)$-dimensional open faces; clearly, $\Gamma \subset \partial\Omega$ with strict inclusion. The equation (2) will be supplemented by boundary conditions. For this purpose, let $\nu(x) = \{\nu_i(x)\}_{i=1}^d$ denote the unit outward normal vector to $\Gamma$ at $x \in \Gamma$. On introducing the *Fichera function* (cf. [48])

$x \in \Gamma \mapsto \beta(x) := (b \cdot \nu)(x) \in \mathbb{R}$, we define the following subsets of $\Gamma$:

$$\Gamma_0 = \{x \in \Gamma : \ \alpha(\nu(x)) > 0\},$$
$$\Gamma_- = \{x \in \Gamma \backslash \Gamma_0 : \ \beta(x) < 0\}, \quad \Gamma_+ = \{x \in \Gamma \backslash \Gamma_0 : \ \beta(x) \geq 0\}.$$

The set $\Gamma_0$ is the *elliptic part* of $\Gamma$, while $\Gamma_- \cup \Gamma_+$ represents the *hyperbolic part* of $\Gamma$. The sets $\Gamma_-$ and $\Gamma_+$ will be referred to as the hyperbolic *inflow* and *outflow* boundary, respectively. Clearly, $\Gamma = \Gamma_0 \cup \Gamma_- \cup \Gamma_+$. If $\Gamma_0 \neq \emptyset$ and has positive $(d-1)$-dimensional Hausdorff measure $\mathcal{H}^{d-1}(\Gamma_0)$, we shall further decompose it into disjoint subsets $\Gamma_D$ and $\Gamma_N$ whose union is $\Gamma_0$, with $\mathcal{H}^{d-1}(\Gamma_D) > 0$. We supplement the partial differential equation (1) with the boundary conditions

$$u = g_D \quad \text{on } \Gamma_D \cup \Gamma_-, \qquad \nu \cdot (a \nabla u) = g_N \quad \text{on } \Gamma_N, \tag{3}$$

and adopt the (physically reasonable) hypothesis that $\beta(x) \geq 0$ for a.e. $x \in \Gamma_N$, whenever $\Gamma_N$ is nonempty. In addition, we assume that the following (standard) positivity hypothesis holds: there exists a positive constant $\hat{c}_0$ such that

$$c(x) + \tfrac{1}{2} \nabla \cdot b(x) \geq \hat{c}_0^2 \quad \text{a.e. } x \in \Omega, \tag{4}$$

and define $c_0 = (c + \frac{1}{2} \nabla \cdot b)^{1/2}$ on $\overline{\Omega}$. Now, consider the following boundary-value problem, corresponding to $g_D = 0$ and $g_N = 0$: find $u$ such that

$$\mathcal{L}u \equiv -\nabla \cdot (a \nabla u) + \nabla \cdot (bu) + cu = f \quad \text{in } \Omega, \tag{5}$$
$$u = 0 \quad \text{on } \Gamma_D \cup \Gamma_-, \tag{6}$$
$$\nu \cdot (a \nabla u) = 0 \quad \text{on } \Gamma_N. \tag{7}$$

**Function spaces and weak formulation.** The classical Sobolev space on $\Omega$ of integer order $m$, $m \geq 0$, will be denoted by $W_q^m(\Omega)$ for $q \in [1, \infty]$; in the case $q = 2$ we write $H^m(\Omega)$ for $W_2^m(\Omega)$; $W_q^0(\Omega)$ is simply $L_q(\Omega)$. $W_p^m(\Omega)$ is equipped with the Sobolev norm $\| \cdot \|_{W_q^m(\Omega)}$ and seminorm $| \cdot |_{W_q^m(\Omega)}$. For the sake of simplicity, we shall write $\| \cdot \|$ instead of $\| \cdot \|_{L_2(\Omega)}$, and $\| \cdot \|_\kappa$ will denote $\| \cdot \|_{L_2(\kappa)}$ for an open subset $\kappa$ of $\Omega$. We let $\mathcal{V} = \{v \in H^1(\Omega) : \gamma_{0,\partial\Omega}(v)|_{\Gamma_D} = 0\}$ where $\gamma_{0,\partial\Omega}(v)$ signifies the trace of $v$ on $\partial\Omega$, and define the inner product $(\cdot, \cdot)_{\mathcal{H}}$ by

$$(w, v)_{\mathcal{H}} := (a \nabla w, \nabla v) + (w, v) + \langle w, v \rangle_{\Gamma_N \cup \Gamma_- \cup \Gamma_+}.$$

Here $(\cdot, \cdot)$ denotes the $L_2$ inner product over $\Omega$ and $\langle w, v \rangle_S = \int_S |\beta| wv \, ds$, with $\beta$ denoting the Fichera function $b \cdot \nu$, as before, and $S \subset \Gamma$. We denote by $\mathcal{H}$ the closure of the space $\mathcal{V}$ in the norm $\| \cdot \|_{\mathcal{H}}$ defined by $\|w\|_{\mathcal{H}} := (w, w)_{\mathcal{H}}^{1/2}$. Clearly, $\mathcal{H}$ is a Hilbert space. For $w \in \mathcal{H}$ and $v \in \mathcal{V}$, we now consider the bilinear form $B(\cdot, \cdot) : \mathcal{H} \times \mathcal{V} \to \mathbb{R}$ defined by

$$B(w, v) := (a \nabla w, \nabla v) - (w, \nabla \cdot (bv)) + (cw, v) + \langle w, v \rangle_{\Gamma_N \cup \Gamma_+},$$

and for $v \in \mathcal{V}$ we introduce the linear functional $L\colon \mathcal{V} \to \mathbb{R}$ by $L(v) := (f, v)$. Note, in particular, that by (4),

$$B(v, v) = (a\nabla v, \nabla v) + \|c_0 v\|^2 + \tfrac{1}{2}\langle v, v\rangle_{\Gamma_N \cup \Gamma_- \cup \Gamma_+} \geq K_0 \|v\|^2_{\mathcal{H}} \quad \text{for all } v \in \mathcal{V},$$

where $K_0 = \min\left(\hat{c}_0^2, \frac{1}{2}\right) > 0$. We shall say that $u \in \mathcal{H}$ is a *weak solution* to the boundary-value problem (5), (6) if

$$B(u, v) = L(v) \quad \text{for all } v \in \mathcal{V}. \tag{8}$$

We note that the boundary conditions $u|_{\Gamma_-} = 0$ on the inflow part $\Gamma_-$ of the hyperbolic boundary $\Gamma \setminus \Gamma_0 = \Gamma_- \cup \Gamma_+$ and the boundary condition $\nu \cdot (a\nabla u) = 0$ on the Neumann part $\Gamma_N$ of the elliptic boundary $\Gamma_0$ are imposed weakly, through (8), while the boundary condition $u|_{\Gamma_D} = 0$ on the Dirichlet part, $\Gamma_D$, of $\Gamma_0$ is imposed strongly, through the choice of the function space $\mathcal{H}$. The existence of a unique weak solution is guaranteed by the following theorem (cf. also Theorem 1.4.1 on p. 29 of [48] and [57] for a similar result in the special case of $\Gamma_N = \emptyset$; for $\Gamma_N \neq \emptyset$ the proof is identical to that in [57]).

**Theorem 2.1.** *Suppose that $c_0(x) \geq \hat{c}_0 > 0$ for all $x \in \overline{\Omega}$. Then, for each $f \in L_2(\Omega)$, there is a unique $u$ in a Hilbert subspace $\hat{\mathcal{H}}$ of $\mathcal{H}$ such that* (8) *holds.*

Next, we shall consider the discretisation of the problem (8), first by a stabilised Galerkin method based on continuous piecewise polynomials, and then using discontinuous piecewise polynomials.

## 3. Continuous piecewise polynomial approximation: the streamline-diffusion method

As in the previous section, we suppose that $\Omega$ is a bounded open polytope in $\mathbb{R}^d$, $d \geq 2$. Let $\mathcal{T}_h = \{\kappa\}$ be an admissible subdivision of $\Omega$ into open element domains $\kappa$ which is subordinate to the decomposition of $\Gamma$ into the subsets $\Gamma_D$, $\Gamma_N$, $\Gamma_-$ and $\Gamma_+$; here $h$ is a piecewise constant mesh function with $h(x) = h_\kappa = \mathrm{diam}(\kappa)$ when $x$ is in element $\kappa \in \mathcal{T}_h$. We shall assume that each $\kappa \in \mathcal{T}_h$ is the image, under a bijective affine map $F_\kappa$, of a fixed *master element* $\hat{\kappa}$, where $\hat{\kappa}$ is either an open unit simplex or an axiparallel open unit hypercube in $\mathbb{R}^d$. We shall also suppose that the family of partitions $\{\mathcal{T}_h\}_{h>0}$ is

(a) *regular* (namely, the closures of any two elements in the subdivision are either disjoint or share a common face of dimension $\leq d-1$); and

(b) *shape-regular* (namely, there exists a positive constant $c_1$, independent of $h$, such that $c_1 h_\kappa^d \leq \mathrm{meas}(\kappa)$ for all $\kappa \in \bigcup_h \mathcal{T}_h$).

For $p \geq 1$, we denote by $\mathcal{P}_p(\hat{\kappa})$ the set of polynomials of degree at most $p$ on $\hat{\kappa}$ when $\hat{\kappa}$ is an open unit simplex; when $\hat{\kappa}$ is an axiparallel open unit hypercube, we let $\mathcal{Q}_p(\hat{\kappa})$ denote the set of all tensor-product polynomials of degree at most $p$ in each coordinate direction. We define the *finite element space*

$$\mathcal{H}_{h,p} = \{v \in \mathcal{H} \cap \mathrm{C}(\overline{\Omega}) : v|_\kappa \in \mathcal{R}_p(\kappa) \text{ for all } \kappa \in \mathcal{T}_h\},$$

where $\mathcal{R}_p(\kappa) = \{w \in \mathrm{L}_1(\kappa) \,:\, w \circ F_\kappa \in \mathcal{R}_p(\hat{\kappa})\}$ and $\mathcal{R}_p$ is either $\mathcal{P}_p$ or $\mathcal{Q}_p$.

Next, we formulate the streamline-diffusion finite element approximation of (8). The method was originally introduced by Hughes and Brooks [36] in 1979 for elliptic transport-dominated diffusion equations. Its analysis was pursued by a number of authors (see [38], [39], [54], for example). The definition of the method stems from the empirical observation that standard Galerkin finite element approximations to transport-dominated diffusion problems exhibit nonphysical numerical oscillations which occur predominantly in the direction of subcharacteristic curves (i.e. the characteristic curves of the underlying hyperbolic problem); the standard Galerkin method is therefore supplemented with numerical diffusion/dissipation in the direction of the subcharacteristics through the inclusion of a *streamline-diffusion stabilisation term*. For a survey of recent perspectives on stabilised and multiscale finite element methods for partial differential equations, including transport-dominated diffusion problems, we refer to the survey paper of Brezzi and Marini [14]; see also [13], [15]. Here, we follow the exposition in [33] and consider the bilinear form $B_\delta(\,\cdot\,,\,\cdot\,)$ defined by

$$\begin{aligned} B_\delta(w, v) = (a\nabla w, \nabla v) - (w, \nabla \cdot (bv)) + (cw, v) \\ + \langle w, v\rangle_{\Gamma_\mathrm{N}\cup\Gamma_+} + \sum_{\kappa\in\mathcal{T}_h} (\hat{\mathcal{L}}w, \delta_\kappa\, b\cdot\nabla v)_\kappa \end{aligned}$$

and the linear functional $\ell_\delta(v) = \sum_\kappa (f,\, v+\delta_\kappa\, b\cdot\nabla v)_\kappa$, where, on element $\kappa \in \mathcal{T}_h$, we define $\hat{\mathcal{L}}$ by $\hat{\mathcal{L}}w = -\nabla\cdot(P_\kappa(a\nabla w)) + b\cdot\nabla w + cw$, $w \in \mathrm{H}^1(\kappa)$, and $P_\kappa$ signifies the orthogonal projection in $[\mathrm{L}_2(\kappa)]^d$ onto $[\mathcal{R}_p(\kappa)]^d$. In these definitions $(\cdot,\cdot)_\kappa$ denotes the $\mathrm{L}_2$ inner product over $\kappa$ and the nonnegative piecewise constant function $\delta$, called the *streamline-diffusion stabilisation parameter*, is defined by $\delta|_\kappa = \delta_\kappa$ for $\kappa \in \mathcal{T}_h$, where $\delta_\kappa$ is a nonnegative constant on element $\kappa$. The precise choice of $\delta$ will be discussed below.

Now, the streamline-diffusion finite element method is defined as follows: find $u_\mathrm{SD} \in \mathcal{H}_{h,p}$ such that

$$B_\delta(u_\mathrm{SD}, v) = \ell_\delta(v) \quad \text{for all } v \in \mathcal{H}_{h,p}. \tag{9}$$

Here, we shall focus on the stability and error analysis of this method. A key property is the following: from (8) and (9) we deduce that if $u \in \mathcal{H} \cap \mathrm{H}^2(\Omega)$ then

$$B_\delta((u - u_\mathrm{SD}), v) = \sum_{\kappa\in\mathcal{T}_h} (\nabla\cdot(a\nabla u - P_\kappa(a\nabla u)), \delta_\kappa b\cdot\nabla v)_\kappa \quad \text{for all } v \in \mathcal{H}_{h,p}. \tag{10}$$

In particular, if $a$ is a constant matrix, then the projection operator $P_\kappa$ can be replaced by the identity operator. In this case the right-hand side in (10) is zero and this identity is then referred to as the *Galerkin orthogonality* property of the streamline-diffusion finite element method (9).

Next we show the stability of the method (9) and state an optimal order *a priori* error bound. The bound will be expressed in terms of the so-called streamline-diffusion norm $|||\cdot|||_{\mathrm{SD}}$ defined by

$$|||v|||_{\mathrm{SD}}^2 = \|\nabla v\|_a^2 + \hat{c}_0^2\|v\|^2 + \|v\|_{\Gamma_{\mathrm{N}}\cup\Gamma_-\cup\Gamma_+}^2 + \|\sqrt{\delta}\, b\cdot\nabla v\|^2,$$

where $\|\nabla v\|_a^2 = (a\nabla v, \nabla v)$. The analysis requires the following results; cf. [33].

**Lemma 3.1** (Inverse inequality). *There exists a positive constant $C_{\mathrm{inv}} = C_{\mathrm{inv}}(c_1)$, independent of $a$, $h_\kappa$ and $p$ such that*

$$\|\nabla\cdot(P_\kappa(a\nabla v))\|_{\mathrm{L}_2(\kappa)} \le C_{\mathrm{inv}}\,\frac{p^2}{h_\kappa}\,\|a\nabla v\|_{\mathrm{L}_2(\kappa)} \quad \textit{for all } v\in\mathcal{H}_{h,p},\ \kappa\in\mathcal{T}_h.$$

**Lemma 3.2.** *Suppose that $M$ is a real $d\times d$ symmetric positive semidefinite matrix and let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^d$; then $|M\xi|^2 \le \rho(M)(M\xi,\xi)$ for all $\xi\in\mathbb{R}^d$, where $\rho(M)=\max_{1\le i\le d}\lambda_i$ is the spectral radius of $M$ and $\lambda_i$, $i=1,\dots,d$, are the (real, nonnegative) eigenvalues of $M$.*

Now we are ready to discuss the coercivity of the bilinear form $B_\delta(\cdot,\cdot)$ over $\mathcal{H}_{h,p}\times\mathcal{H}_{h,p}$. To this end we define $\mathcal{T}_h' = \{\kappa\in\mathcal{T}_h : \|b\|_{\mathrm{L}_\infty(\kappa)}\neq 0\}$.

**Proposition 3.3.** *Suppose that the streamline-diffusion parameter $\delta_\kappa$ on element $\kappa$ is selected, with the convention $1/0=\infty$, so that*

$$0\le\delta_\kappa\le\tfrac{1}{2}\min\left(\frac{h_\kappa^2}{(C_{\mathrm{inv}})^2\,\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)}\,p^4},\ \frac{\hat{c}_0^2}{\|c\|_{\mathrm{L}_\infty(\kappa)}^2}\right) \quad \textit{for all } \kappa\in\mathcal{T}_h'. \tag{11}$$

*Then, the bilinear form $B_\delta(\cdot,\cdot)$ is coercive on $\mathcal{H}_{h,p}\times\mathcal{H}_{h,p}$, i.e.*

$$B_\delta(v,v)\ge\tfrac{1}{2}|||v|||_{\mathrm{SD}}^2 \quad \textit{for all } v\in\mathcal{H}_{h,p}.$$

*Proof.* Integrating by parts gives

$$\begin{aligned} B_\delta(v,v) \ge{}& (a\nabla v,\nabla v) + \int_\Omega c_0^2 v^2\,\mathrm{d}x + \tfrac{1}{2}\|v\|_{\Gamma_{\mathrm{N}}\cup\Gamma_-\cup\Gamma_+}^2 + \|\sqrt{\delta}\,b\cdot\nabla v\|^2 \\ &- \sum_{\kappa\in\mathcal{T}_h'}\delta_\kappa\left(\|\nabla\cdot(P_\kappa(a\nabla v))\|_\kappa + \|cv\|_\kappa\right)\|b\cdot\nabla v\|_\kappa, \end{aligned} \tag{12}$$

for $v\in\mathcal{H}_{h,p}$. Now, using Lemma 3.1 and Lemma 3.2 with $M=a$ and $\xi=\nabla v$,

$$\|\nabla\cdot(P_\kappa(a\nabla v))\|_\kappa + \|cv\|_\kappa \le C_{\mathrm{inv}}\,\frac{p^2}{h_\kappa}\,\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)}^{1/2}\|\nabla v\|_{a,\kappa} + \|c\|_{\mathrm{L}_\infty(\kappa)}\|v\|_\kappa,$$

with the notation $\|\nabla v\|_{a,\kappa} = (a\nabla v, \nabla v)_\kappa^{1/2}$. Thus, for any real number $\gamma > 0$,

$$\sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \left(\|\nabla\cdot(P_\kappa(a\nabla v))\|_\kappa + \|cv\|_\kappa\right)\|b\cdot\nabla v\|_\kappa \le \gamma \sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \|b\cdot\nabla v\|_\kappa^2$$
$$+ \frac{1}{2\gamma}\sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \|c\|_{\mathrm{L}_\infty(\kappa)}^2 \|v\|_\kappa^2 + \frac{1}{2\gamma}\sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \,(C_{\mathrm{inv}})^2 \,\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)} \frac{p^4}{h_\kappa^2}\,\|\nabla v\|_{a,\kappa}^2.$$

Choosing $\gamma = 1/2$, we deduce from (12) and the definition of $c_0$ that

$$B_\delta(v,v) \ge \|\nabla v\|_a^2 + \hat{c}_0^2\|v\|^2 + \tfrac{1}{2}\|v\|_{\Gamma_{\mathrm{N}}\cup\Gamma_-\cup\Gamma_+}^2 + \tfrac{1}{2}\|\sqrt{\delta}\; b\cdot\nabla v\|^2$$
$$- \sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \|c\|_{\mathrm{L}_\infty(\kappa)}^2\|v\|_\kappa^2 - \sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \,(C_{\mathrm{inv}})^2 \,\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)} \frac{p^4}{h_\kappa^2}\,\|\nabla v\|_{a,\kappa}^2.$$

Selecting the streamline-diffusion parameter as in (11), the result follows. □

**Corollary 3.4** (Stability). *Under the hypotheses of Proposition* 3.3,

$$|||u_{\mathrm{SD}}|||_{\mathrm{SD}}^2 \le 4\hat{c}_0^{-2}\|f\|^2 + 4\sum_{\kappa\in\mathcal{T}_h'} \delta_\kappa \|f\|_\kappa^2.$$

In particular, if $f = 0$ then $u_{\mathrm{SD}} = 0$; since $\mathcal{H}_{h,p}$ is a finite-dimensional linear space, it follows that (9) has a unique solution $u_{\mathrm{SD}} \in \mathcal{H}_{h,p}$ for any $f \in \mathrm{L}_2(\Omega)$. The next result concerns the accuracy of the method (9).

**Theorem 3.5.** *Let the streamline-diffusion parameter $\delta_\kappa$ be chosen so that*

$$0 < \delta_\kappa \le \frac{1}{2}\min\left(\frac{h_\kappa^2}{(C_{\mathrm{inv}})^2\,\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)}\,p^4}\,,\ \frac{\hat{c}_0^2}{\|c\|_{\mathrm{L}_\infty(\kappa)}^2}\right) \quad \textit{for all } \kappa \in \mathcal{T}_h',$$

*with the convention $1/0 = \infty$. Then, assuming that $u \in \mathcal{H}\cap \mathrm{H}^k(\Omega)\cap \mathrm{C}(\overline{\Omega})$, with a positive integer $k$ and $a \in [\mathrm{W}_\infty^{k-1}(\kappa)]^{d\times d}$, $\kappa \in \mathcal{T}_h$, the following error bound holds:*

$$|||u - u_{\mathrm{SD}}|||_{\mathrm{SD}} \le C\Big(\sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2\tau-1}}{p^{2k-1}} M_\kappa(a,b,c,h_\kappa,p)\|u\|_{\mathrm{H}^k(\kappa)}^2\Big)^{1/2},$$

*where $\tau = \min(p+1,k)$, $C$ is a positive constant which depends only on $c_1$ and $k$, $M_\kappa(a,b,c,h_\kappa,p) = A_\kappa(p/h_\kappa) + B_\kappa + C_\kappa(h_\kappa/p)$, with*

$$A_\kappa = \begin{cases} \|a\|_{\mathrm{L}_\infty(\kappa)} + \dfrac{\|a\|_{\mathrm{L}_\infty(\kappa)}^2 + \|a\|_{\mathrm{W}_\infty^{k-1}(\kappa)}^2}{\|\rho(a)\|_{\mathrm{L}_\infty(\kappa)}}, & \textit{when } \|\rho(a)\|_{\mathrm{L}_\infty(\kappa)} \ne 0,\\ 0, & \textit{otherwise},\end{cases}$$

$$B_\kappa = \begin{cases} \|b\|_{\mathrm{L}_\infty(\kappa)}(1 + D_\kappa + D_\kappa^{-1}), & \textit{when } \|b\|_{\mathrm{L}_\infty(\kappa)} \ne 0,\\ 0, & \textit{otherwise},\end{cases}$$

$$C_\kappa = 1 + c_0 + c_0^{-2}\,\|c - \nabla\cdot b\|_{\mathrm{L}_\infty(\kappa)}^2, \quad D_\kappa = \delta_\kappa\|b\|_{\mathrm{L}_\infty(\kappa)}p/h_\kappa.$$

Theorem 3.5 is an extension of classical *a priori* error bounds for the streamline-diffusion discretisation of a first-order hyperbolic problem and a second-order elliptic transport-diffusion problem with $a$ isotropic and constant; see, for example, [37], [54] ($h$-version) and [29] ($hp$-version), and [33] for a proof in the case of $\Gamma_N = \emptyset$.

## 4. Discontinuous piecewise polynomial approximation: the discontinuous Galerkin method

Discontinuous Galerkin finite element methods (DGFEMs, for short) date back to the early 1970s; they were simultaneously proposed by Reed & Hill [52] in 1973 for the numerical solution of the neutron transport equation and by Nitsche [45] in 1971 as a nonstandard scheme for the approximation of second-order elliptic equations. Since the early 1970s there has been extensive work on the development of these methods for a wide range of applications; for an excellent historical survey of DGFEMs up until 2000 we refer to the paper of Cockburn, Karniadakis and Shu in the volume [19].

One of the key advantages of the DGFEM in comparison with standard Galerkin finite element methods based on continuous piecewise polynomials, such as the streamline-diffusion finite method discussed in the previous section, is their high degree of locality: the computational stencil of the DGFEM remains very compact even as the degree of the approximating polynomial is increased. Hence, high-order adaptive $hp$- and spectral element approximations may be handled in a particularly flexible and simple manner. Indeed, $hp$-adaptive DGFEMs offer tremendous gains in terms of computational efficiency in comparison with standard mesh refinement algorithms which only incorporate local $h$-refinement with a given (fixed) polynomial degree. For discussions concerning various $hp$-refinement strategies see [1], [8], [32], [34], [53], [58]. A further attractive property of the discontinuous Galerkin finite element method for a transport-dominated diffusion problem is that, unlike its counterpart based on continuous piecewise polynomials, the method is stable even in the absence of streamline-diffusion stabilisation.

In this section, we survey *a priori* and *a posteriori* error bounds for discontinuous Galerkin finite element approximations of second-order partial differential equations with nonnegative characteristic form. We shall then show how the *a posteriori* error bound can be used to drive an $hp$-adaptive finite element algorithm. The presentation in this section is based on the paper [25].

We consider shape-regular meshes $\mathcal{T}_h = \{\kappa\}$ that partition the domain $\Omega$ into open element domains $\kappa$, with possible *hanging nodes*. We shall suppose that the mesh is 1-irregular in the sense that there is at most one hanging node per $(d-1)$-dimensional element-face, e.g. the barycenter of the face. We denote by $h$ the piecewise constant mesh function with $h(x) \equiv h_\kappa = \mathrm{diam}(\kappa)$ when $x$ is in element $\kappa$. Let each $\kappa \in \mathcal{T}_h$ be a smooth bijective image of a fixed master element $\hat{\kappa}$, that is, $\kappa = F_\kappa(\hat{\kappa})$ for all $\kappa \in \mathcal{T}_h$, where $\hat{\kappa}$ is either the open unit simplex $\hat{\kappa}_S = \{\hat{x} = (\hat{x}_1, \ldots, \hat{x}_d) \in \mathbb{R}^d : 0 < x_1 + \cdots + x_d < 1,\ x_i > 0,\ i = 1, \ldots, d\}$, or the open hypercube

$\hat{\kappa}_{\mathrm{C}} = (-1, 1)^d$ in $\mathbb{R}^d$. On $\hat{\kappa}$ we define spaces of polynomials of degree $p \geq 1$ as follows: $\mathcal{Q}_p = \mathrm{span}\{\hat{x}^\alpha : 0 \leq \alpha_i \leq p,\ 1 \leq i \leq d\}$, $\mathcal{P}_p = \mathrm{span}\{\hat{x}^\alpha : 0 \leq |\alpha| \leq p\}$. To each $\kappa \in \mathcal{T}_h$ we assign an integer $p_\kappa \geq 1$; collecting the $p_\kappa$ and $F_\kappa$ in the vectors $\boldsymbol{p} = \{p_\kappa : \kappa \in \mathcal{T}_h\}$ and $\boldsymbol{F} = \{F_\kappa : \kappa \in \mathcal{T}_h\}$, respectively, we introduce the finite element space

$$S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}) = \{u \in \mathrm{L}_2(\Omega) : u|_\kappa \circ F_\kappa \in \mathcal{Q}_{p_\kappa} \text{ if } F_\kappa^{-1}(\kappa) = \hat{\kappa}_{\mathrm{C}} \\ \text{and } u|_\kappa \circ F_\kappa \in \mathcal{P}_{p_\kappa} \text{ if } F_\kappa^{-1}(\kappa) = \hat{\kappa}_{\mathrm{S}};\ \kappa \in \mathcal{T}_h\}.$$

We assign to $\mathcal{T}_h$ the *broken Sobolev space* of composite order $\boldsymbol{s} = \{s_\kappa : \kappa \in \mathcal{T}_h\}$ defined by $\mathrm{H}^{\boldsymbol{s}}(\Omega, \mathcal{T}_h) = \{u \in \mathrm{L}_2(\Omega) : u|_\kappa \in \mathrm{H}^{s_\kappa}(\kappa) \text{ for all } \kappa \in \mathcal{T}_h\}$, equipped with the *broken Sobolev norm*

$$\|u\|_{\boldsymbol{s}, \mathcal{T}_h} = \Big( \sum_{\kappa \in \mathcal{T}_h} \|u\|^2_{\mathrm{H}^{s_\kappa}(\kappa)} \Big)^{1/2}.$$

When $s_\kappa = s$ for all $\kappa \in \mathcal{T}_h$, we write $\mathrm{H}^s(\Omega, \mathcal{T}_h)$ and $\|u\|_{s, \mathcal{T}_h}$.

An *interior face* of $\mathcal{T}_h$ is defined as the (non-empty) $(d-1)$-dimensional interior of $\partial\kappa_i \cap \partial\kappa_j$, where $\kappa_i$ and $\kappa_j$ are two adjacent elements of $\mathcal{T}_h$, not necessarily matching. A *boundary face* of $\mathcal{T}_h$ is defined as the (non-empty) $(d-1)$-dimensional interior of $\partial\kappa \cap \Gamma$, where $\kappa$ is a boundary element of $\mathcal{T}_h$. We denote by $\Gamma_{\mathrm{int}}$ the union of all interior faces of $\mathcal{T}_h$. Given a face $e \subset \Gamma_{\mathrm{int}}$, shared by the two elements $\kappa_i$ and $\kappa_j$, where the indices $i$ and $j$ satisfy $i > j$, we write $\nu_e$ to denote the (numbering-dependent) unit normal vector which points from $\kappa_i$ to $\kappa_j$; on boundary faces we put $\nu_e = \nu$. Further, for $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)$ we define the jump of $v$ across $e$ and the mean value of $v$ on $e$, respectively, by $[v] = v|_{\partial\kappa_i \cap e} - v|_{\partial\kappa_j \cap e}$ and $\langle v \rangle = \frac{1}{2}(v|_{\partial\kappa_i \cap e} + v|_{\partial\kappa_j \cap e})$. On a boundary face $e \subset \partial\kappa$, we set $[v] = v|_{\partial\kappa \cap e}$ and $\langle v \rangle = v|_{\partial\kappa \cap e}$. Finally, given a function $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)$ and an element $\kappa \in \mathcal{T}_h$, we denote by $v_\kappa^+$ (respectively, $v_\kappa^-$) the interior (respectively, exterior) trace of $v$ defined on $\partial\kappa$ (respectively, $\partial\kappa \backslash \Gamma$). Since below it will always be clear from the context which element $\kappa$ in the subdivision $\mathcal{T}_h$ the quantities $v_\kappa^+$ and $v_\kappa^-$ correspond to, for the sake of notational simplicity we shall suppress the letter $\kappa$ in the subscript and write, respectively, $v^+$ and $v^-$ instead. Given that $\kappa$ is an element in the subdivision $\mathcal{T}_h$, we denote by $\partial\kappa$ the union of $(d-1)$-dimensional open faces of $\kappa$. Let $x \in \partial\kappa$ and suppose that $\nu_\kappa(x)$ denotes the unit outward normal vector to $\partial\kappa$ at $x$. With these conventions, we define the inflow and outflow parts of $\partial\kappa$, respectively, by $\partial_-\kappa = \{x \in \partial\kappa : b(x) \cdot \nu_\kappa(x) < 0\}$, $\partial_+\kappa = \{x \in \partial\kappa : b(x) \cdot \nu_\kappa(x) \geq 0\}$.

For simplicity of presentation, we suppose that each entry of the matrix $a$ is piecewise continuous on $\mathcal{T}_h$ and belongs to $S^{\boldsymbol{0}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$. With minor changes only, our results can easily be extended to the case when each entry of $\sqrt{a}$ belongs to $S^{\boldsymbol{q}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$, where the composite polynomial degree vector $\boldsymbol{q}$ has nonnegative entries; for more general $a$, see [23]. In the following, we write $\bar{a} = |\sqrt{a}\,|_2^2$, where $|\cdot|_2$ denotes the matrix norm subordinate to the $l_2$-vector norm on $\mathbb{R}^d$ and $\bar{a}_\kappa = \bar{a}|_\kappa$; by $\bar{a}_{\tilde{\kappa}}$ we denote the arithmetic mean of the values $\bar{a}_{\kappa'}$ over those elements $\kappa'$ (including $\kappa$ itself) that share a $(d-1)$-dimensional face with $\kappa$.

The $hp$-DGFEM approximation of (1), (3) is defined as follows: find $u_{\mathrm{DG}}$ in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ such that

$$B_{\mathrm{DG}}(u_{\mathrm{DG}}, v) = \ell_{\mathrm{DG}}(v) \quad \text{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}). \tag{13}$$

Here, the bilinear form $B_{\mathrm{DG}}(\,\cdot\,,\cdot\,)$ is defined by

$$B_{\mathrm{DG}}(w, v) = B_a(w, v) + B_b(w, v) + \theta B_e(v, w) - B_e(w, v) + B_\sigma(w, v)\,,$$

where

$$\begin{aligned}
B_a(w, v) &= \sum_{\kappa\in\mathcal{T}_h} \int_\kappa a\nabla w \cdot \nabla v \,\mathrm{d}x,\\
B_b(w, v) &= \sum_{\kappa\in\mathcal{T}_h} \Big\{ -\int_\kappa (w\, b\cdot\nabla v - cwv)\,\mathrm{d}x + \int_{\partial_+\kappa} (b\cdot\nu_\kappa)\, w^+ v^+ \,\mathrm{d}s\\
&\qquad\qquad + \int_{\partial_-\kappa\setminus\Gamma} (b\cdot\nu_\kappa)\, w^- v^+ \,\mathrm{d}s\Big\},\\
B_e(w, v) &= \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \langle (a\nabla w)\cdot\nu_e\rangle [v]\,\mathrm{d}s,\\
B_\sigma(w, v) &= \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma [w][v]\,\mathrm{d}s,
\end{aligned}$$

and the linear functional $\ell_{\mathrm{DG}}(\,\cdot\,)$ is given by

$$\begin{aligned}
\ell_{\mathrm{DG}}(v) = \sum_{\kappa\in\mathcal{T}_h} \Big\{ &\int_\kappa f v\,\mathrm{d}x - \int_{\partial_-\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)} (b\cdot\nu_\kappa)\, g_{\mathrm{D}}\, v^+ \,\mathrm{d}s\\
&+ \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} \theta\, g_{\mathrm{D}}((a\nabla v^+)\cdot\nu_\kappa)\,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v^+ \,\mathrm{d}s\\
&+ \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} \sigma g_{\mathrm{D}} v^+ \,\mathrm{d}s\Big\}.
\end{aligned}$$

Here, $\sigma$ is defined by

$$\sigma|_e = C_\sigma \frac{\langle \bar{a} p^2\rangle}{\langle h\rangle} \quad \text{for } e \subset \Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}, \tag{14}$$

where $C_\sigma$ is a positive constant, called the *discontinuity-penalisation* parameter; cf. [30]. We shall adopt the convention that edges $e \subset \Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}$ with $\sigma|_e = 0$ are omitted from the integrals appearing in the definition of $B_\sigma(w, v)$ and $\ell_{\mathrm{DG}}(v)$, although we shall not highlight this explicitly in our notation; the same convention is adopted in the case of integrals where the integrand contains the factor $1/\sigma$. Thus, in particular, the definition of the DG-norm, cf. (15) below, is meaningful even if $\sigma|_e$ happens to be equal to zero on certain edges $e \subset \Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}$, given that such edges are understood to be excluded from the region of integration.

Selecting the parameter $\theta = 1$ gives rise to the so-called *Nonsymmetric Interior Penalty* (*NIP*) *method*, while setting $\theta = -1$ yields the *Symmetric Interior Penalty* (*SIP*) *scheme*; in the following we write SIP/NIP to denote the symmetric/nonsymmetric versions of the interior penalty method.

While a symmetric discretisation of a symmetric differential operator seems quite natural, the NIP scheme is often preferred, especially for transport-dominated problems where the underlying discretisation matrix is nonsymmetric anyway, as it is stable for any choice of the parameter $C_\sigma > 0$; see, for example, [2], [30], [51], and Theorem 4.1 below. On the other hand, the SIP scheme is only stable when $C_\sigma > 0$ is chosen sufficiently large. In terms of accuracy, both schemes converge at the optimal rate when the error is measured in terms of the DG-norm (cf. (15) below), but the lack of adjoint consistency (see, [2]) of the NIP method leads to suboptimal convergence of the error when measured in terms of the $L_2$ norm. In this case, the SIP scheme is still optimally convergent, while the NIP method is suboptimal by a full order; however, numerical experiments indicate that in practice the $L_2$ norm of the error arising from the NIP scheme converges to zero at the optimal rate when the polynomial degree $p$ is odd, cf. [30]. Thereby, in practice the loss of optimality of the NIP scheme when the error is measured in terms of the $L_2$ norm only arises for even $p$. However, we showed in [25] that, for $p \geq 2$, the lack of adjoint consistency of the NIP scheme leads to an even more dramatic deterioration of its convergence rate when the error is measured in terms of a certain (linear) target functional $J(\cdot)$, such as $J: v \mapsto \int_\Omega v(x)\psi(x)\,\mathrm{d}x$, for example, where $\psi$ is a given weight-function: for fixed $p$ the error measured in terms of $J(\cdot)$ behaves like $\mathcal{O}(h^{2p})$ when the SIP scheme is employed, while for the NIP scheme, in general we only have the rate of convergence $\mathcal{O}(h^p)$ as $h$ tends to zero. For related work on *a posteriori* error estimation for DGFEMs with interior penalty, see e.g. Becker *et al.* [9], [10] and Rivière & Wheeler [53]. For further perspectives on the construction and postprocessing of DGFEMs, see [12], [18].

Before embarking on the analysis of the discontinuous Galerkin method (13), we define the DG-norm $|||\cdot|||_{\mathrm{DG}}$ by

$$
\begin{aligned}
|||w|||_{\mathrm{DG}}^2 = \sum_{\kappa\in\mathcal{T}_h} \Big( & \|\nabla w\|_{a,\kappa}^2 + \|c_0 w\|_\kappa^2 + \frac{1}{2}\|w^+\|_{\partial_-\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)}^2 \\
& + \frac{1}{2}\|w^+\|_{\partial_+\kappa\cap\Gamma}^2 + \frac{1}{2}\|w^+ - w^-\|_{\partial_-\kappa\setminus\Gamma}^2 \Big) \\
& + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma[w]^2\,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \frac{1}{\sigma}\langle (a\,\nabla w)\cdot \nu_e\rangle^2\,\mathrm{d}s,
\end{aligned}
\tag{15}
$$

where $\|\nabla w\|_{a,\kappa}^2 = (a\nabla w, \nabla w)_\kappa$, $\|\cdot\|_\tau$, $\tau \subset \partial\kappa$, denotes the (semi)norm induced by the (semi)inner-product $(v, w)_\tau = \int_\tau |b\cdot\nu_\kappa| v w\,\mathrm{d}s$, and $c_0$ is as defined in (4). The above definition of $|||\cdot|||_{\mathrm{DG}}$ represents a slight modification of the norm considered in [30]; for $a > 0$, $b \equiv 0$, (15) corresponds to the norm proposed by Baumann *et al.* [8], [47] and Baker *et al.* [5], cf. [51]. With this notation, we state the following coercivity result for the bilinear form $B_{\mathrm{DG}}(\cdot,\cdot)$ over $S^p(\Omega, \mathcal{T}_h, \boldsymbol{F}) \times S^p(\Omega, \mathcal{T}_h, \boldsymbol{F})$.

**Theorem 4.1.** *With $\sigma$ defined as in* (14)*, there exists a positive constant $C$, which depends only on the dimension $d$ and the shape-regularity of $\mathcal{T}_h$, such that*

$$B_{\mathrm{DG}}(v, v) \geq C|||v|||_{\mathrm{DG}}^2 \quad \textit{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}),$$

*provided that the constant $C_\sigma$ arising in the definition of the discontinuity penalisation parameter $\sigma$ is chosen so that $C_\sigma > 0$ arbitrary when $\theta = 1$, and $C_\sigma \geq C'_\sigma$ with $C'_\sigma$ sufficiently large when $\theta \neq 1$.*

This result is an extension of the coercivity result derived by Prudhomme *et al.* [51] with $b \equiv 0$; see also [30] for the proof in the case when $\theta = 1$. For the case of $a \equiv 0$, the connection of stabilisation to upwinding has been discussed in [16]. In particular, Theorem 4.1 implies that (13) has a unique solution for any $f \in \mathrm{L}_2(\Omega)$, any $g_{\mathrm{D}} \in \mathrm{L}_2(\Gamma_{\mathrm{D}})$ and any $g_{\mathrm{N}} \in \mathrm{L}_2(\Gamma_{\mathrm{N}})$. Theorem 4.1 also indicates that while the NIP scheme is coercive over $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}) \times S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ for any choice of the constant $C_\sigma > 0$ arising in the definition of the discontinuity-penalisation parameter $\sigma$, the SIP scheme (corresponding to $\theta = -1$) is only coercive if $C_\sigma$ is chosen sufficiently large; see [25] for details about the minimum size of $C_\sigma$.

Henceforth, we shall assume that the solution $u$ to the boundary value problem (1), (3) is sufficiently smooth: namely, $u \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ and that $u$ and $(a\nabla u) \cdot \nu_e$ are continuous across each face $e \subset \partial\kappa \backslash \Gamma$ that intersects the subdomain of ellipticity, $\Omega_a = \{x \in \overline{\Omega} : \boldsymbol{\zeta}^\top a(x)\boldsymbol{\zeta} > 0 \text{ for all } \boldsymbol{\zeta} \in \mathbb{R}^d\}$. If this smoothness requirement is violated, the discretisation method has to be modified accordingly, cf. [30]. We note that under these assumptions, the following Galerkin orthogonality property holds:

$$B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, v) = 0 \quad \text{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}). \tag{16}$$

We shall assume that $b \in [\mathrm{W}^1_\infty(\Omega)]^d$ is such that

$$b \cdot \nabla_{\mathcal{T}_h} v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}, \boldsymbol{F}) \quad \text{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}, \boldsymbol{F}). \tag{17}$$

Let us denote by $\Pi_p$ the orthogonal projector in $\mathrm{L}_2(\Omega)$ onto the finite element space $S^{\boldsymbol{p}}(\Omega, \mathcal{T}, \boldsymbol{F})$. We remark that this choice of projector is essential in the following *a priori* error analysis, in order to ensure that $(u - \Pi_p u, b \cdot \nabla_{\mathcal{T}_h} v) = 0$ for all $v$ in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}, \boldsymbol{F})$. If, on the other hand, the scheme (13) is supplemented by streamline-diffusion stabilisation, then alternative choices of $\Pi_p$ may be employed (see [29], [59], for example); in that case, hypothesis (17) is redundant. We now decompose the global error $u - u_{\mathrm{DG}}$ as

$$u - u_{\mathrm{DG}} = (u - \Pi_p u) + (\Pi_p u - u_{\mathrm{DG}}) \equiv \eta + \xi. \tag{18}$$

**Lemma 4.2.** *Assume that* (4) *and* (17) *hold and let $\beta_1|_\kappa = \|c/c_0^2\|_{\mathrm{L}_\infty(\kappa)}$; then there exists a positive constant $C$ that depends only on $d$ and the shape-regularity of $\mathcal{T}_h$*

*such that the functions $\xi$ and $\eta$ defined by* (18) *satisfy the following inequality*

$$|||\xi|||_{\mathrm{DG}}^2 \le C\Big\{ \sum_{\kappa\in\mathcal{T}_h} \big(\|\sqrt{a}\,\nabla\eta\|_\kappa^2 + \beta_1^2\|c_0\eta\|_\kappa^2 + \|\eta^+\|_{\partial_+\kappa\cap\Gamma}^2 + \|\eta^-\|_{\partial_-\kappa\setminus\Gamma}^2\big) + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \frac{1}{\sigma}\langle(a\nabla\eta)\cdot\nu_e\rangle^2\,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma[\eta]^2\,\mathrm{d}s\Big\}.$$

The proof is given in [25]. We also need the following result concerning the approximation properties of the projector $\Pi_p$; for simplicity, we restrict ourselves to 1-irregular, shape-regular meshes consisting of affine equivalent $d$-parallelepiped elements (cf. [4], [30], and also [24] for similar results in augmented Sobolev spaces).

**Lemma 4.3.** *Suppose that $\kappa \in \mathcal{T}_h$ is a $d$-parallelepiped of diameter $h_\kappa$ and that $u|_\kappa \in \mathrm{H}^{k_\kappa}(\kappa)$, $k_\kappa \ge 0$, for $\kappa \in \mathcal{T}_h$. Then, the following approximation results hold:*

$$\|u-\Pi_p u\|_{\mathrm{L}_2(\kappa)} \le C\frac{h_\kappa^{s_\kappa}}{p_\kappa^{k_\kappa}}\|u\|_{\mathrm{H}^{k_\kappa}(\kappa)}, \quad \|u-\Pi_p u\|_{\mathrm{L}_2(\partial\kappa)} \le C\frac{h_\kappa^{s_\kappa-1/2}}{p_\kappa^{k_\kappa-1/2}}\|u\|_{\mathrm{H}^{k_\kappa}(\kappa)},$$

$$|u-\Pi_p u|_{\mathrm{H}^1(\kappa)} \le C\frac{h_\kappa^{s_\kappa-1}}{p_\kappa^{k_\kappa-3/2}}\|u\|_{\mathrm{H}^{k_\kappa}(\kappa)}, \quad |u-\Pi_p u|_{\mathrm{H}^1(\partial\kappa)} \le C\frac{h_\kappa^{s_\kappa-3/2}}{p_\kappa^{k_\kappa-5/2}}\|u\|_{\mathrm{H}^{k_\kappa}(\kappa)},$$

*where $1 \le s_\kappa \le \min(p_\kappa+1, k_\kappa)$ and $C$ is a constant independent of $u$, $h_\kappa$ and $p_\kappa$, but dependent on the dimension $d$ and the shape-regularity of $\mathcal{T}_h$.*

For the rest of this section, we assume that the polynomial degree vector $\boldsymbol{p}$, with $p_\kappa \ge 1$, $\kappa \in \mathcal{T}_h$, has *bounded local variation*; *i.e.*, there exists a constant $\rho \ge 1$ such that, for any pair of elements $\kappa$ and $\kappa'$ which share a $(d-1)$-dimensional face,

$$\rho^{-1} \le p_\kappa/p_{\kappa'} \le \rho. \tag{19}$$

On noting that $\eta = u - \Pi_p u$ and combining Lemmas 4.2 and 4.3, we deduce that

$$|||\xi|||_{\mathrm{DG}}^2 \le C\sum_{\kappa\in\mathcal{T}_h}\left(\alpha\frac{h_\kappa^{2(s_\kappa-1)}}{p_\kappa^{2(k_\kappa-3/2)}} + \beta_2\frac{h_\kappa^{2s_\kappa}}{p_\kappa^{2k_\kappa}} + \gamma\frac{h_\kappa^{2(s_\kappa-1/2)}}{p_\kappa^{2(k_\kappa-1/2)}}\right)\|u\|_{\mathrm{H}^{k_\kappa}(\kappa)}^2,$$

where $\alpha|_\kappa = \bar{a}_{\tilde{\kappa}}$, $\beta_2|_\kappa = (\beta_1|_\kappa)^2\,\|c_0\|_{\mathrm{L}_\infty(\kappa)}^2$, $(\beta_1|_\kappa = \|c/(c_0)^2\|_{\mathrm{L}_\infty(\kappa)}$, cf. Lemma 4.2), $\gamma|_\kappa = \|b\|_{\mathrm{L}_\infty(\kappa)}$ and $C$ is a positive constant that depends only on $d$, the parameter $\rho$ in (19) and the shape-regularity of $\mathcal{T}_h$. The DG-norm $|||\eta|||_{\mathrm{DG}}$ of $\eta = u - \Pi_p u$ can be estimated directly using Lemma 4.3 to show that a bound analogous to that on $|||\xi|||_{\mathrm{DG}}$ above holds. Hence, a bound on the discretisation error $u - u_{\mathrm{DG}} = \xi + \eta$ in the DG-norm $|||\cdot|||_{\mathrm{DG}}$ is obtained via the triangle inequality (see [30] for details).

Very often in practice the objective of the computation is not the approximation of the analytical solution $u$ in a given norm, but controlling the error in an output/target-functional $J(\cdot)$ of the solution. Relevant examples of output functionals include the lift and drag coefficients of a body immersed into a viscous fluid, the local mean value

of the field, or its flux through the outflow boundary of the computational domain. Here we give a brief survey of *a posteriori* and *a priori* error bounds for general linear target functionals $J(\cdot)$ of the solution; for related work, we refer to [11], [26], [28], [31], [32], [34], [42], [58], [59], for example, and to the recent monograph of Bangerth & Rannacher [6]. We shall confine ourselves to Type I (dual-weighted) *a posteriori* bounds; the computationally simpler, but cruder, Type II error bounds will not be discussed here (see Giles & Süli [22]).

**Type I *a posteriori* error analysis.** We proceed as in [34], [58] and begin by considering the following *dual* or *adjoint* problem: find $z \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ such that

$$B_{\mathrm{DG}}(w, z) = J(w) \quad \text{for all } w \in \mathrm{H}^2(\Omega, \mathcal{T}_h). \tag{20}$$

Let us assume that (20) possesses a unique solution. Clearly, the validity of this assumption depends, *inter alia*, on the choice of the linear functional under consideration. We shall return to this issue below; see also [25], [28], [34].

For a given linear functional $J(\cdot)$ the *a posteriori* error bound will be expressed in terms of the finite element residual $R_{\mathrm{int}}$ defined on $\kappa \in \mathcal{T}_h$ by $R_{\mathrm{int}}|_\kappa = (f - \mathcal{L}u_{\mathrm{DG}})|_\kappa$, which measures the extent to which $u_{\mathrm{DG}}$ fails to satisfy the differential equation on the union of the elements $\kappa$ in the mesh $\mathcal{T}_h$; thus we refer to $R_{\mathrm{int}}$ as the *internal residual*. Also, since $u_{\mathrm{DG}}$ only satisfies the boundary conditions approximately, the differences $g_{\mathrm{D}} - u_{\mathrm{DG}}$ and $g_{\mathrm{N}} - (a\nabla u_{\mathrm{DG}}) \cdot \nu$ are not necessarily zero on $\Gamma_{\mathrm{D}} \cup \Gamma_-$ and $\Gamma_{\mathrm{N}}$, respectively; thus we define the *boundary residuals* $R_{\mathrm{D}}$ and $R_{\mathrm{N}}$ by $R_{\mathrm{D}}|_{\partial\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)} = (g_{\mathrm{D}} - u^+_{\mathrm{DG}})|_{\partial\kappa\cap(\Gamma_{\mathrm{D}}\cup\Gamma_-)}$ and $R_{\mathrm{N}}|_{\partial\kappa\cap\Gamma_{\mathrm{N}}} = (g_{\mathrm{N}} - (a\nabla u^+_{\mathrm{DG}}) \cdot \nu)|_{\partial\kappa\cap\Gamma_{\mathrm{N}}}$, respectively. By using the divergence theorem, the Galerkin orthogonality property (16) may be rewritten as follows:

$$\begin{aligned}
0 = B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, v) &= \ell_{\mathrm{DG}}(v) - B_{\mathrm{DG}}(u_{\mathrm{DG}}, v) \\
&= \sum_{\kappa\in\mathcal{T}_h} \Bigg\{ \int_\kappa R_{\mathrm{int}} v \,\mathrm{d}x - \int_{\partial_-\kappa\cap\Gamma} (b\cdot\nu_\kappa)\, R_{\mathrm{D}} v^+ \,\mathrm{d}s \\
&\qquad + \int_{\partial_-\kappa\setminus\Gamma} (b\cdot\nu_\kappa)\,[u_{\mathrm{DG}}]v^+ \,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} \theta\, R_{\mathrm{D}}((a\nabla v^+)\cdot\nu_\kappa)\,\mathrm{d}s \\
&\qquad + \int_{\partial\kappa\cap\Gamma_{\mathrm{D}}} \sigma R_{\mathrm{D}} v^+ \,\mathrm{d}s + \int_{\partial\kappa\cap\Gamma_{\mathrm{N}}} R_{\mathrm{N}} v^+ \,\mathrm{d}s \\
&\qquad - \int_{\partial\kappa\setminus\Gamma} \left( \frac{\theta}{2}\,[u_{\mathrm{DG}}](a\nabla v^+)\cdot\nu_\kappa + \frac{1}{2}[(a\nabla u_{\mathrm{DG}})\cdot\nu_\kappa]v^+ \right) \mathrm{d}s \\
&\qquad - \int_{\partial\kappa\setminus\Gamma} \sigma [u_{\mathrm{DG}}] v^+ \,\mathrm{d}s \Bigg\}
\end{aligned} \tag{21}$$

for all $v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$. The starting point is the following result from [25].

**Theorem 4.4.** *Let $u$ and $u_{\mathrm{DG}}$ denote the solutions of* (1), (3) *and* (13), *respectively, and suppose that the dual solution $z$ is defined by* (20). *Then, the following error*

*representation formula holds:*

$$J(u) - J(u_{\mathrm{DG}}) = \mathcal{E}_{\Omega}(u_{\mathrm{DG}}, h, p, z - z_{h,p}) \equiv \sum_{\kappa \in \mathcal{T}_h} \eta_{\kappa}, \tag{22}$$

*where*

$$\begin{aligned}
\eta_{\kappa} = & \int_{\kappa} R_{\mathrm{int}}(z - z_{h,p})\,\mathrm{d}x - \int_{\partial_{-}\kappa \cap \Gamma} (b \cdot \nu_{\kappa})\, R_{\mathrm{D}}(z - z_{h,p})^{+}\,\mathrm{d}s \\
& + \int_{\partial_{-}\kappa \setminus \Gamma} (b \cdot \nu_{\kappa})\, [u_{\mathrm{DG}}](z - z_{h,p})^{+}\,\mathrm{d}s \\
& + \int_{\partial\kappa \cap \Gamma_{\mathrm{D}}} \theta\, R_{\mathrm{D}}((a\nabla(z - z_{h,p})^{+}) \cdot \nu_{\kappa})\,\mathrm{d}s \\
& + \int_{\partial\kappa \cap \Gamma_{\mathrm{D}}} \sigma\, R_{\mathrm{D}}(z - z_{h,p})^{+}\,\mathrm{d}s + \int_{\partial\kappa \cap \Gamma_{\mathrm{N}}} R_{\mathrm{N}}(z - z_{h,p})^{+}\,\mathrm{d}s \\
& - \int_{\partial\kappa \setminus \Gamma} \left\{ \frac{\theta}{2}\, [u_{\mathrm{DG}}](a\nabla(z - z_{h,p})^{+}) \cdot \nu_{\kappa} + \frac{1}{2}[(a\nabla u_{\mathrm{DG}}) \cdot \nu_{\kappa}](z - z_{h,p})^{+} \right\} \mathrm{d}s \\
& - \int_{\partial\kappa \setminus \Gamma} \sigma [u_{\mathrm{DG}}](z - z_{h,p})^{+}\,\mathrm{d}s
\end{aligned} \tag{23}$$

*for all* $z_{h,p} \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$.

*Proof.* On choosing $w = u - u_{\mathrm{DG}}$ in (20) and recalling the linearity of $J(\cdot)$ and the Galerkin orthogonality property (21), we deduce that

$$\begin{aligned}
J(u) - J(u_{\mathrm{DG}}) &= J(u - u_{\mathrm{DG}}) = B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, z) \\
&= B_{\mathrm{DG}}(u - u_{\mathrm{DG}}, z - z_{h,p}),
\end{aligned} \tag{24}$$

and hence (22), with $\eta_{\kappa}$ defined by (23), using the definitions of the residuals. □

**Corollary 4.5.** *Under the assumptions of Theorem* 4.4*, and with* $\eta_{\kappa}$ *defined as in* (23)*, the following Type I* a posteriori *error bound holds:*

$$|J(u) - J(u_{\mathrm{DG}})| \leq \mathcal{E}_{|\Omega|}(u_{\mathrm{DG}}, h, p, z - z_{h,p}) \equiv \sum_{\kappa \in \mathcal{T}_h} |\eta_{\kappa}|\,. \tag{25}$$

As discussed in [6], [11], [26], [58], the local *weighting terms* involving the difference between the dual solution $z$ and its projection/interpolant $z_{h,p}$ onto $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ appearing in the Type I bound (25) contain useful information concerning the global transport of the error. Therefore, we shall retain the weighting terms involving the (unknown) dual solution $z$ in our bound and approximate them numerically, — instead of eliminating $z$, as one would in the derivation of a structurally simpler, but cruder, Type II *a posteriori* bound. However, before proceeding any further, we need to consider more carefully the dual problem defined by (20). Let us suppose, for example,

that the aim of the computation is to approximate the (weighted) mean value of the solution $u$; *i.e.*, $J(\cdot) \equiv M_\psi(\cdot)$, where $M_\psi(w) = \int_\Omega w\psi \, \mathrm{d}x$ and $\psi \in \mathrm{L}_2(\Omega)$. When $\theta = -1$, performing integration by parts, we find that the corresponding dual solution $z = z^{\mathrm{SIP}}$ is the solution of the following mesh-dependent problem: find $z$ such that

$$\begin{aligned} \mathcal{L}^* z \equiv -\nabla \cdot (a\nabla z) - b \cdot \nabla z + cz = \psi \quad & \text{in } \Omega, \\ z = 0 \quad & \text{on } \Gamma_{\mathrm{D}} \cup \Gamma_+, \\ (b \cdot \nu)z + (a\nabla z) \cdot \nu = 0 \quad & \text{on } \Gamma_{\mathrm{N}}. \end{aligned}$$

Thus, for $\theta = -1$ the dual problem is well-posed for this choice of target functional. We remark that, since in this case the dual problem formed by transposing the arguments in the bilinear form $B_{\mathrm{DG}}(\cdot,\cdot) = B^{\mathrm{SIP}}_{\mathrm{DG}}(\cdot,\cdot)$ involves the formal adjoint of the partial differential operator $\mathcal{L}$, $B^{\mathrm{SIP}}_{\mathrm{DG}}(\cdot,\cdot)$ is said to be *adjoint consistent*, cf. Arnold *et al.* [2]; in particular, when $\theta = -1$ and the primal and dual solutions are sufficiently smooth, the error in the functional will be seen to exhibit an optimal order of convergence. As we shall explain below by means of *a priori* error analysis, the situation is dramatically different when $\theta \neq -1$: then, the bilinear form $B_{\mathrm{DG}}(\cdot,\cdot)$ is *not* adjoint consistent; this, in turn, leads to degradation of the convergence rate of the error in the computed functional $J(\cdot)$ as the finite element space $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ is enriched (by reducing $h$ or by increasing the polynomial degree vector $\boldsymbol{p}$). Once again, we refer to [25] for technical details.

***A priori* error bounds.** We continue to use the superscripts SIP and NIP to distinguish between the two methods and write $B^{\mathrm{SIP}}_{\mathrm{DG}}(\cdot,\cdot) \equiv B_{\mathrm{DG}}(\cdot,\cdot)$ when $\theta = -1$ and $B^{\mathrm{NIP}}_{\mathrm{DG}}(\cdot,\cdot) \equiv B_{\mathrm{DG}}(\cdot,\cdot)$ when $\theta = 1$. The corresponding numerical solutions $u^{\mathrm{SIP}}_{\mathrm{DG}}$ and $u^{\mathrm{NIP}}_{\mathrm{DG}}$ satisfy the following problems: find $u^{\mathrm{SIP}}_{\mathrm{DG}}$ in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ such that

$$B^{\mathrm{SIP}}_{\mathrm{DG}}(u^{\mathrm{SIP}}_{\mathrm{DG}}, v) = \ell_{\mathrm{DG}}(v) \quad \text{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F});$$

and find $u^{\mathrm{NIP}}_{\mathrm{DG}}$ in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ such that

$$B^{\mathrm{NIP}}_{\mathrm{DG}}(u^{\mathrm{NIP}}_{\mathrm{DG}}, v) = \ell_{\mathrm{DG}}(v) \quad \text{for all } v \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F}),$$

respectively. The starting point for the *a priori* error analysis is the identity (24) in the proof of Theorem 4.4. Again, using the above notation, we see that

$$J(u) - J(u^{\mathrm{SIP}}_{\mathrm{DG}}) = B^{\mathrm{SIP}}_{\mathrm{DG}}(u - u^{\mathrm{SIP}}_{\mathrm{DG}}, z^{\mathrm{SIP}} - z_{h,p})$$

when the SIP scheme is employed, while for the NIP scheme, we have

$$J(u) - J(u^{\mathrm{NIP}}_{\mathrm{DG}}) = B^{\mathrm{NIP}}_{\mathrm{DG}}(u - u^{\mathrm{NIP}}_{\mathrm{DG}}, z^{\mathrm{NIP}} - z_{h,p})$$

for all $z_{h,p}$ in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$. Here, $z^{\mathrm{SIP}}$ and $z^{\mathrm{NIP}}$ are the analytical solutions to the following dual problems: find $z^{\mathrm{SIP}} \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ such that

$$B^{\mathrm{SIP}}_{\mathrm{DG}}(w, z^{\mathrm{SIP}}) = J(w) \quad \text{for all } w \in \mathrm{H}^2(\Omega, \mathcal{T}_h);$$

and find $z^{\text{NIP}} \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ such that

$$B_{\text{DG}}^{\text{NIP}}(w, z^{\text{NIP}}) = J(w) \quad \text{for all } w \in \mathrm{H}^2(\Omega, \mathcal{T}_h).$$

Hence, for all $z_{h,p} \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$,

$$J(u) - J(u_{\text{DG}}) = B_{\text{DG}}(u - u_{\text{DG}}, z^{\text{SIP}} - z_{h,p}) - (1+\theta) B_e(z^{\text{SIP}}, u - u_{\text{DG}}^{\text{NIP}}), \quad (26)$$

where $u_{\text{DG}}$ is either $u_{\text{DG}}^{\text{SIP}}$ or $u_{\text{DG}}^{\text{NIP}}$, depending on whether $\theta = -1$ or $\theta = 1$, respectively. In particular, the second term on the right-hand side of (26) is absent for the SIP scheme, i.e. when $\theta = -1$, but it is present when the NIP scheme is employed, *i.e.*, when $\theta = 1$. Since this second term is of lower order than the first term in (26), it will lead to suboptimal rates of convergence as the finite element space $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ is enriched in the case of $\theta \neq -1$.

**Theorem 4.6.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral domain, $\mathcal{T}_h = \{\kappa\}$ a shape-regular subdivision of $\Omega$ into $d$-parallelepipeds and $\boldsymbol{p}$ a polynomial degree vector of bounded local variation. Let* (4) *and* (17) *hold, let the entries of $a$ be piecewise constant on $\mathcal{T}_h$, and $u|_\kappa \in \mathrm{H}^{k_\kappa}(\kappa)$, $k_\kappa \geq 2$, for $\kappa \in \mathcal{T}_h$, $z^{\text{SIP}}|_\kappa \in \mathrm{H}^{l_\kappa}(\kappa)$, $l_\kappa \geq 2$, for $\kappa \in \mathcal{T}_h$; then, the solution $u_{\text{DG}} \in S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ of* (13) *satisfies the error bound*

$$|J(u) - J(u_{\text{DG}})|^2 \leq C \left\{ \sum_{\kappa \in \mathcal{T}_h} \left( \alpha \frac{h_\kappa^{2(s_\kappa - 1)}}{p_\kappa^{2(k_\kappa - 3/2)}} + \beta_3 \frac{h_\kappa^{2s_\kappa}}{p_\kappa^{2k_\kappa}} + \gamma \frac{h_\kappa^{2(s_\kappa - 1/2)}}{p_\kappa^{2(k_\kappa - 1/2)}} \right) \|u\|^2_{\mathrm{H}^{k_\kappa}(\kappa)} \right\}$$

$$\times \left\{ \sum_{\kappa \in \mathcal{T}_h} \left( \alpha \frac{h_\kappa^{2(t_\kappa - 1)}}{p_\kappa^{2(l_\kappa - 3/2)}} + \beta_4 \frac{h_\kappa^{2t_\kappa}}{p_\kappa^{2l_\kappa}} + \gamma \frac{h_\kappa^{2(t_\kappa - 1/2)}}{p_\kappa^{2(l_\kappa - 1)}} \right) \|z^{\text{SIP}}\|^2_{\mathrm{H}^{l_\kappa}(\kappa)} + (1+\theta) \|z^{\text{SIP}}\|^2_{2,\mathcal{T}_h} \right\}$$

*for $1 \leq s_\kappa \leq \min(p_\kappa + 1, k_\kappa)$, $1 \leq t_\kappa \leq \min(p_\kappa + 1, l_\kappa)$, $p_\kappa \geq 1$, $\kappa \in \mathcal{T}_h$, where $\alpha|_\kappa = \bar{a}_{\tilde{\kappa}}$, $\beta_3|_\kappa = (1 + (\beta_1|_\kappa)^2)\|c_0\|^2_{\mathrm{L}_\infty(\kappa)}$, $(\beta_1|_\kappa = \|c(x)/(c_0(x))^2\|_{\mathrm{L}_\infty(\kappa)})$, $\beta_4|_\kappa = \|(c + \nabla \cdot b)/c_0\|^2_{\mathrm{L}_\infty(\kappa)}$, $\gamma|_\kappa = \|b\|_{\mathrm{L}_\infty(\kappa)}$ and $C$ is a constant depending on the dimension $d$, the parameter $\rho$ from* (19) *and the shape-regularity of $\mathcal{T}_h$.*

If we assume uniform orders $p_\kappa = p$, $s_\kappa = s$, $t_\kappa = t$, $k_\kappa = k$, $l_\kappa = l$, where $s$, $t$, $k$ and $l$ are positive integers, and $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$, then, in the diffusion-dominated case (viz. $b \approx 0$), Theorem 4.6, with $\theta = -1$ implies that for the SIP scheme

$$|J(u) - J(u_{\text{DG}})| \leq C\, (h^{s+t-2}/p^{k+l-2})\, p\, \|u\|_{\mathrm{H}^k(\Omega)} \|z^{\text{SIP}}\|_{\mathrm{H}^l(\Omega)}, \quad (27)$$

where $1 \leq s \leq \min(p+1, k)$ and $1 \leq t \leq \min(p+1, l)$. This error bound is optimal with respect to $h$ and suboptimal in $p$ by a full order. We note, however, that 'order-doubling' of the rate of convergence in $|J(u) - J(u_{\text{DG}})|$ observed when the SIP scheme is employed, as expressed by (27), is lost when the NIP method is used. In the hyperbolic case ($a \equiv 0$), the bound in Theorem 4.6 becomes

$$|J(u) - J(u_{\text{DG}})| \quad \leq \quad C\, (h^{s+t-1}/p^{k+l-1})\, p^{1/2}\, \|u\|_{\mathrm{H}^k(\Omega)} \|z^{\text{SIP}}\|_{\mathrm{H}^l(\Omega)}.$$

This error bound is optimal in $h$ and suboptimal in $p$ by $p^{1/2}$ (cf. also [34]).

**Adaptive algorithm.** In the light of Theorem 4.6, we now confine ourselves to the SIP scheme ($\theta = -1$). For a user-defined tolerance TOL, we consider the problem of designing an $hp$-finite element space $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ such that the inequality $|J(u) - J(u_{\mathrm{DG}})| \leq \mathtt{TOL}$ holds, subject to the constraint that the number of degrees of freedom in $S^{\boldsymbol{p}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ is minimized. Following [34], we use the *a posteriori* error bound (25) with $z$ replaced by a discontinuous Galerkin approximation $\hat{z}$ computed on the same mesh $\mathcal{T}_h$ as for the primal solution $u_{\mathrm{DG}}$, but with a higher degree polynomial, *i.e.*, $\hat{z} \in S^{\hat{\boldsymbol{p}}}(\Omega, \mathcal{T}_h, \boldsymbol{F})$ , $\hat{\boldsymbol{p}} = \boldsymbol{p} + \boldsymbol{p}_{\mathrm{inc}}$; in our experiments we set $\boldsymbol{p}_{\mathrm{inc}} = \mathbf{1}$, cf. [26], [32], [58]. Thereby, in practice we enforce the stopping criterion

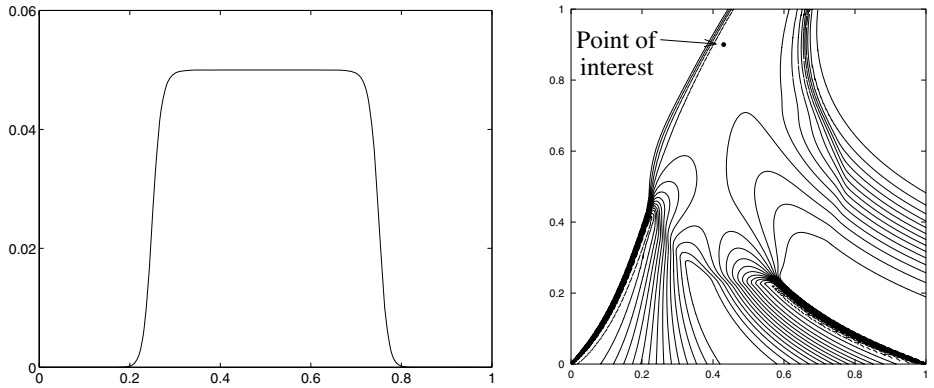
$$\hat{\mathcal{E}}_{|\Omega|} \equiv \mathcal{E}_{|\Omega|}(u_{\mathrm{DG}}, h, p, \hat{z} - z_{h,p}) \leq \mathtt{TOL}\,. \tag{28}$$

If (28) is not satisfied, then the elements are marked for refinement/derefinement according to the size of the (approximate) error indicators $|\hat{\eta}_\kappa|$; these are defined analogously to $|\eta_\kappa|$ in (23) with $z$ replaced by $\hat{z}$. In our experiments we use the fixed fraction mesh refinement algorithm, with refinement and derefinement fractions set to 20% and 10%, respectively.

Once an element $\kappa \in \mathcal{T}_h$ has been flagged for refinement or derefinement, a decision must be made whether the local mesh size $h_\kappa$ or the local degree $p_\kappa$ of the approximating polynomial should be altered. The choice to perform either $h$-refinement/derefinement or $p$-refinement/derefinement is based on the local smoothness of the primal and dual solutions $u$ and $z$, respectively; cf. [32], [34]. Let us first consider the case when an element has been flagged for *refinement*. If $u$ or $z$ are locally smooth, then $p$-refinement will be more effective than $h$-refinement, since the error will be expected to decay quickly within the current element $\kappa$ as $p_\kappa$ is increased. On the other hand, if both $u$ and $z$ have low regularity within the element $\kappa$, then $h$-refinement will be performed. To ensure that the desired level of accuracy is achieved efficiently, in [34] an automatic procedure was developed for deciding when to $h$- or $p$-refine, based on the smoothness-estimation strategy proposed by Ainsworth & Senior [1]. For a review of various $hp$-adaptive strategies as well as descriptions of new algorithms based on Sobolev index estimation *via* local Legendre expansions, we refer to [31], [32]. If an element has been flagged for *derefinement*, then the strategy implemented here is to coarsen the mesh in low-error-regions where either the primal or dual solutions $u$ and $z$, respectively, are smooth and decrease the degree of the approximating polynomial in low-error-regions when both $u$ and $z$ are insufficiently regular, cf. [34].

**Numerical experiments.** We explore the performance of the $hp$-adaptive strategy outlined above for the symmetric version of the interior penalty method, applied to a mixed hyperbolic-elliptic problem with discontinuous boundary data (cf. [25]). We let $a = \varepsilon(x)I$, where $\varepsilon = \frac{1}{2}\delta(1 - \tanh((r - 1/4)(r + 1/4)/\gamma))$, $r^2 = (x - 1/2)^2 + (y - 1/2)^2$ and $\delta \geq 0$ and $\gamma > 0$ are constants. Let $b = (2y^2 - 4x + 1, 1 + y)$, $c = -\nabla \cdot b$ and $f = 0$. With $\delta > 0$ and $0 < \gamma \ll 1$, the diffusion parameter $\varepsilon$

is approximately equal to $\delta$ in the circular region defined by $r < 1/4$, where the underlying partial differential equation is uniformly elliptic. In this example, we set $\delta = 0.05$ and $\gamma = 0.01$; a cross-section of $\varepsilon$ along $0 \le x \le 1$, $y = 1/2$ is shown in Figure 1. As the value of $r$ is increased beyond $1/4$, the function $\varepsilon$ rapidly decreases through a layer of width $\mathcal{O}(\gamma)$; for example, when $r > 0.336$ we have $\varepsilon < 10^{-15}$, so from the computational point of view $\varepsilon$ is zero to within rounding error; in this




(a) (b)

Figure 1. (a) Profile of $\varepsilon$ along $y = 0.5$, $0 \le x \le 1$; (b) DGFEM approximation to the primal problem on a $129 \times 129$ mesh with piecewise bilinear elements ($\boldsymbol{p} = \mathbf{1}$); from [25].

region, the partial differential equation undergoes a change of type becoming, in effect, hyperbolic. Thus we shall refer to the part of $\Omega$ with $r > 1/4 + \mathcal{O}(\gamma)$ as the *hyperbolic region*, while the set of points in $\Omega$ with $r \le 1/4$ will be called the *elliptic region*; of course, strictly speaking, the partial differential equation is elliptic in the whole of $\overline{\Omega}$. The characteristics associated with the hyperbolic part of the operator enter the computational domain $\Omega$ from three sides of $\Gamma$, namely through the vertical edges placed along $x = 0$ and $x = 1$ and the horizontal edge along $y = 0$; the characteristics exit $\Omega$ through the horizontal edge along $y = 1$. On the union of these three faces we prescribe the following boundary condition:

$$u(x, y) = \begin{cases} 1 & \text{for } x = 0 \ , \ 0 < y \le 1, \\ \sin^2(\pi x) & \text{for } 0 \le x \le 1 \ , \ y = 0, \\ \mathrm{e}^{-50y^4} & \text{for } x = 1 \, , \ 0 < y \le 1. \end{cases}$$

Figure 1 shows the numerical approximation to (1) using the SIP method on a uniform $129 \times 129$ uniform square mesh with $\boldsymbol{p} = \mathbf{1}$. Let us suppose that the objective of the computation is to calculate the value of the analytical solution $u$ at a certain point of interest, $x = (0.43, 0.9)$, *i.e.*, $J(u) = u(0.43, 0.9)$; cf. Figure 1. The true value of the functional is given by $J(u) = 0.704611313375$.

In Table 1 we show the performance of our adaptive finite element algorithm using $hp$-refinement. Clearly, the computed Type I *a posteriori* error bound (25) is very sharp in the sense that it overestimates the true error in the computed functional by a factor of about 1–8 only, and by a factor of only 3.34 on average on the meshes that arise in the course of our adaptive $hp$-refinement.

Table 1. History of the adaptive $hp$-refinement. The effectivity index is defined as the ratio of the *a posteriori* error bound $\sum_\kappa |\hat{\eta}_\kappa|$ and the error $|J(u) - J(u_{\mathrm{DG}})|$; from [25].

| Nodes | Elements | Degrees of freedom | $\lvert J(u) - J(u_{\mathrm{DG}})\rvert$ | $\sum_\kappa \lvert\hat{\eta}_\kappa\rvert$ | Effectivity index |
|---|---|---|---|---|---|
| 81 | 64 | 576 | 1.924e-02 | 3.330e-02 | 1.73 |
| 99 | 76 | 740 | 1.056e-02 | 1.085e-02 | 1.03 |
| 162 | 130 | 1451 | 1.006e-02 | 2.290e-02 | 2.28 |
| 241 | 193 | 2483 | 7.400e-04 | 2.385e-03 | 3.22 |
| 302 | 244 | 3776 | 3.760e-05 | 2.754e-04 | 7.32 |
| 323 | 262 | 4777 | 1.270e-05 | 1.026e-04 | 8.08 |
| 396 | 325 | 6916 | 9.896e-06 | 2.245e-05 | 2.27 |
| 487 | 403 | 9941 | 1.224e-06 | 6.466e-06 | 5.28 |
| 577 | 481 | 13528 | 4.656e-07 | 1.163e-06 | 2.50 |
| 713 | 601 | 19855 | 2.449e-07 | 2.582e-07 | 1.05 |
| 960 | 820 | 31019 | 1.574e-08 | 3.202e-08 | 2.03 |
| 1313 | 1132 | 47406 | 6.531e-10 | 2.154e-09 | 3.30 |

Figure 2 shows $|J(u) - J(u_{\mathrm{DG}})|$, using both $h$- and $hp$-refinement, against the square-root of the number of degrees of freedom on a linear-log scale. After the initial transient, the error in the computed functional using $hp$-refinement is seen to become (on average) a straight line, which indicates exponential convergence of $J(u_{\mathrm{DG}})$ to $J(u)$; this occurs since $z^{\mathrm{SIP}}$ is a real analytic function in the regions of the computational domain where $u$ is not smooth and *vice versa*. Figure 2 also demonstrates the superiority of the adaptive $hp$-refinement strategy over the standard adaptive $h$-refinement algorithm when $\mathtt{TOL} \lessapprox 10^{-3}$.

On the final mesh the error between $J(u)$ and $J(u_{\mathrm{DG}})$ using $hp$-refinement is over 4 orders of magnitude smaller than the corresponding quantity when $h$-refinement is used alone.

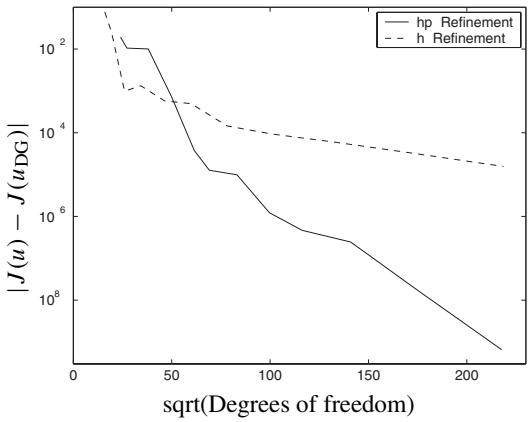



Figure 2. Comparison between $h$- and $hp$-adaptive mesh refinement; from [25].

Figure 3 depicts the primal mesh after 11 adaptive mesh refinement steps. We display the $h$-mesh alone, as well as the corresponding distribution of the polynomial degree on this mesh and the percentage of elements with that degree. We see that some $h$-refinement of the primal mesh has taken place in the region of the computational

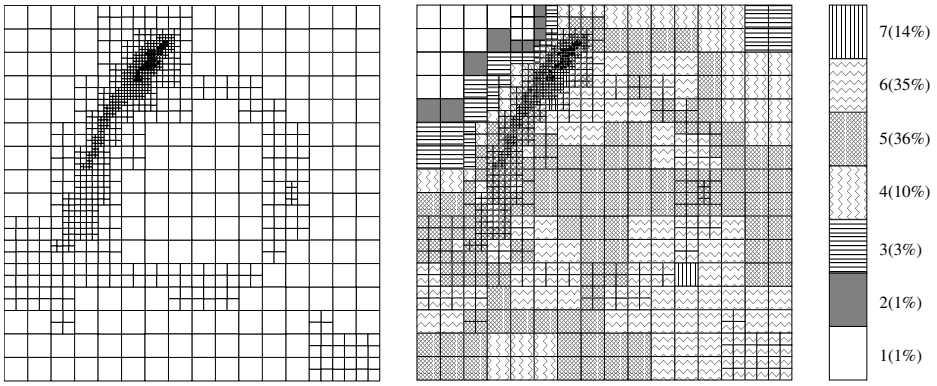



Figure 3. $h$- and $hp$-meshes after 11 refinements, with 1313 nodes, 1132 elements and 47406 degrees of freedom: here, $|J(u) - J(u_{\mathrm{DG}})| = 6.531 \times 10^{-10}$; from [25].

domain upstream of the point of interest, as well as in the circular region where the underlying partial differential equation changes type. Once the $h$-mesh has adequately captured the structure of the primal and dual solutions, the $hp$-adaptive algorithm performs $p$-refinement elsewhere in the domain of dependence of the point of interest.

## 5. High-dimensional transport-diffusion problems

We conclude by giving some pointers to recent results on stabilised sparse finite element methods for high-dimensional partial differential equations (1) with nonnegative characteristic form (cf. [57]). Such high-dimensional equations arise from a number of important applications in physics, chemistry, biology and finance. The origins of sparse tensor-product constructions and hyperbolic cross spaces can be traced back to Babenko [3] and Smolyak [56]; we refer to the papers of Temlyakov [61], DeVore, Konyagin & Temlyakov [20] for the study of high-dimensional approximation problems, to the works of Wasilkowski & Woźniakowski [62] and Novak & Ritter [46] for high-dimensional integration problems and associated complexity questions, to the paper of Zenger [63] for an early contribution to the numerical solution of high-dimensional elliptic equations, to the articles by von Petersdorff & Schwab [50] and Hoang & Schwab [27] for the analysis of sparse-grid methods for high-dimensional elliptic multiscale problems and parabolic equations, respectively, and to the recent survey article of Bungartz & Griebel [17].

Suppose that $\Omega = (0,1)^d$, $\Gamma_{\rm N} = \emptyset$ and $g_{\rm D} = 0$ in (3), and that the operator $\mathcal{L}$ in (1) has constant coefficients. In the simplest case, the construction of the finite element space $\hat{V}_0^L \subset \mathcal{H}$ begins by taking the tensor product of $d$ copies of a finite element space of univariate hierarchical continuous piecewise linear functions ($p = 1$) on a uniform mesh of size $h_L = 2^{-L}$, $L \geq 1$. The resulting tensor-product space $V_0^L$ has dimension $\dim(V_0^L) = \mathcal{O}(h_L^{-d})$. Clearly, the use of this space would lead to exponential growth of computational complexity for fixed $h_L$, as $d$ increases. Thus, the idea is to reduce the complexity of the computation for large $d$ by sparsifying the space $V_0^L$; the resulting sparse finite element space is denoted $\hat{V}_0^L$ and has only $\dim(\hat{V}_0^L) = \mathcal{O}(h_L^{-1}|\log h_L|^{d-1})$ degrees of freedom. The relevant result from [57], stated in the theorem below, is that, with a careful choice of the streamline-diffusion stabilisation parameter $\delta_L$ and assuming that $u \in \mathcal{H}^2(\Omega) \cap \mathcal{H}$, where $\mathcal{H}^2(\Omega) = \{v : D^\alpha v \in {\rm L}_2(\Omega),\ |\alpha|_\infty \leq 2\}$ is the space of functions with $\rm L_2$-bounded mixed second derivatives, one can ensure that this reduction of computational complexity is achieved at essentially no loss in accuracy in the streamline-diffusion finite element method compared to the case when the full tensor-product space $V_0^L$ is used instead of the sparse space $\hat{V}_0^L$.

**Theorem 5.1.** *Let $f \in {\rm L}_2(\Omega)$, $c > 0$ and $u \in \mathcal{H}^2(\Omega) \cap \mathcal{H}$. Then, the following bound holds for the error $u - u_{\rm SD}$ between the analytical solution $u$ of* (8) *and its sparse finite element approximation $u_{\rm SD} \in \hat{V}_0^L$, with $L \geq 1$ and $h = h_L = 2^{-L}$:*

$$
\begin{aligned}
&|||u - u_{\rm SD}|||_{\rm SD}^2 \\
&\qquad \leq C(u)\left\{|a|h_L^2 + h_L^4|\log_2 h_L|^{2(d-1)}\max\left(\frac{|a|}{h_L^2},\ \frac{d|b|}{h_L|\log_2 h_L|^{d-1}},\ c\right)\right\}
\end{aligned}
$$

*with the streamline-diffusion parameter $\delta_L$ defined by the formula*

$$\delta_L := K_\delta \min\left(\frac{h_L^2}{|a|}, \frac{h_L|\log_2 h_L|^{d-1}}{d|b|}, \frac{1}{c}\right),$$

*with $K_\delta \in \mathbb{R}_{>0}$ a constant, independent of $h_L$ and $d$, and $C(u) = \text{Const.}\|u\|^2_{\mathcal{H}^2(\Omega)}$ where* Const. *is a positive constant independent of the discretisation parameter $h_L$.*

We refer to [57] for further technical details, including the proof of this result.

## 6. Concluding remarks

We surveyed continuous stabilised and discontinuous Galerkin finite element methods for the numerical solution of second-order partial differential equations with nonnegative characteristic form. We stated *a priori* and residual-based *a posteriori* error bounds, and in the case of the discontinuous Galerkin method we showed how the *a posteriori* bound may be used to drive an *hp*-adaptive finite element algorithm. We also commented on the use of sparse stabilised finite element methods for high-dimensional transport-dominated diffusion equations: stochastic analysis and kinetic theory are particularly fertile sources of Fokker–Planck equations of this kind [41]. The numerical solution of high-dimensional partial differential equations has been an active area of research in recent years [17], though the bulk of the research has been confined to self-adjoint elliptic and parabolic equations. As we have briefly indicated, extensions of these results to the, vastly richer, class of partial differential equations with nonnegative characteristic form are feasible, and we expect that activities in this direction will continue to flourish.

**Acknowledgments.** I am grateful to Franco Brezzi, Bernardo Cockburn, Kathryn Gillow, Paul Houston, Donatella Marini, Rolf Rannacher and Christoph Schwab for numerous stimulating discussions on the ideas presented in this paper. The computational experiments in Section 4 were performed by Paul Houston.

## References

[1] Ainsworth, M., Senior, B., An adaptive refinement strategy for *hp*-finite element computations. *Appl. Numer. Math.* **26** (1998), 165–178.

[2] Arnold, D. N., Brezzi, F., Cockburn, B., Marini, D., Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39** (2002), 1749–1779.

[3] Babenko, K., Approximation by trigonometric polynomials is a certain class of periodic functions of several variables. *Soviet Math. Dokl.* **1** (1960), 672–675; Russian original in *Dokl. Akad. Nauk SSSR* **132** (1960), 982–985.

[4] Babuška, I., Suri, M., The *hp*-version of the finite element method with quasi-uniform meshes. *RAIRO Modél. Math. Anal. Numér.* **21** (1987), 199–238.

[5] Baker, G. A., Jureidini, W. N., Karakashian, O.A., Piecewise solenoidal vector fields and the Stokes problem. *SIAM J. Numer. Anal.* **27** (1990), 1466–1485.

[6] Bangerth, W., Rannacher, R., *Adaptive Finite Element Methods for Solving Differential Equations.* Lectures in Mathematics ETH Zürich. Birkhäuser, Basel 2003.

[7] Barrett, J. W., Schwab, C., Süli, E., Existence of global weak solutions for some polymeric flow models. *Math. Models Methods Appl. Sci.* **6** (15) (2005), 939–983.

[8] Baumann, C., An *hp*-adaptive discontinuous Galerkin FEM for computational fluid dynamics. Doctoral Dissertation, TICAM, UT Austin, Texas, 1997.

[9] Becker, R., Hansbo, P., Discontinuous Galerkin methods for convection-diffusion problems with arbitrary Péclet number. In *Numerical Mathematics and Advanced Applications: Proceedings of the 3rd European Conference* (P. Neittaanmäki, T. Tiihonen and P. Tarvainen, eds.), World Scientific, River Edge, NJ, 2000, 100–109.

[10] Becker, R., Hansbo, P., Larson, M. G., Energy norm a posteriori error estimation for discontinuous Galerkin methods. Chalmers Finite Element Center Preprint 2001-11, Chalmers University of Technology, Sweden, 2001.

[11] Becker, R., Rannacher, R., Weighted A Posteriori Error Control in FE Methods. Preprint 1, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, Heidelberg, Germany, 1996.

[12] Brezzi, F., Cockburn, B., Marini, L. D., Süli, E., Stabilization mechanisms in discontinuous Galerkin finite element methods. *Comput. Methods Appl. Mech. Engrg.*, to appear.

[13] Brezzi, F., Hughes, T. J. R., Marini, L. D., Russo, A., Süli, E., A priori error analysis of residual-free bubbles for advection-diffusion problems. *SIAM J. Numer. Anal.* **36** (1999), 1939–1948 (electronic).

[14] Brezzi, F, Marini, L. D., Subgrid phenomena and numerical schemes. In *Mathematical modeling and numerical simulation in continuum mechanics* (Yamaguchi, 2000), Lecture Notes Comput. Sci. Eng. 19, Springer-Verlag, Berlin 2002, 73–89.

[15] Brezzi, F., Marini, L. D., Süli, E., Residual-free bubbles for advection-diffusion problems: the general error analysis. *Numer. Math.* **85** (2000), 31–47.

[16] Brezzi, F., Marini, L. D., Süli, E., Discontinuous Galerkin methods for first-order hyperbolic problems. *Math. Models Methods Appl. Sci.* **14** (12) (2004), 1893–1903.

[17] Bungartz, H.-J., Griebel, M., Sparse grids. *Acta Numerica* **13** (2004), 1–123.

[18] Cockburn, B., Luskin, M., Shu, C.-W., Süli, E., Postprocessing of the discontinuous Galerkin finite element method. *Math. Comp.* **72** (2003), 577-606.

[19] Cockburn, B., Karniadakis, G. E., Shu, C.-W., The development of discontinuous Galerkin methods. In *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. E. Karniadakis and C.-W. Shu, eds.), Lecture Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin 2000, 3–50.

[20] DeVore, R., Konyagin, S., Temlyakov, V., Hyperbolic wavelet approximation. *Constr. Approx.* **14** (1998), 1–26.

[21] Elf, J., Lötstedt, P., Sjöberg, P., Problems of high dimension in molecular biology. In *Proceedings of the* 19th *GAMM-Seminar Leipzig* (W. Hackbusch, ed.), 2003, 21–30.

[22] Giles, M. B., Süli, E., Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica* **11** (2002), 145–236.

[23] Georgoulis, E. H., Lasis, A., A note on the design of hp-version interior penalty discontinuous Galerkin finite element methods for degenerate problems. Advance Access published online on October 4, 2005, *IMA Journal of Numerical Analysis*.

[24] Georgoulis, E. H., Süli, E., Optimal error estimates for the $h$-version interior penalty discontinuous Galerkin finite element method. *IMA Journal of Numerical Analysis* **25** (2005), 205–220.

[25] Harriman, K., Houston, P., Senior, B., Süli, E., $hp$-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. In *Recent Advances in Scientific Computing and Partial Differential Equations* (C.-W. Shu, T. Tang, and S.-Y. Cheng, eds.), Contemp. Math. 330, Amer. Math. Soc., Providence, RI, 2003, 89–119.

[26] Hartmann, R., Houston, P., Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM J. Sci. Comp.* **24** (2002), 979–1004.

[27] Hoang, V. H., Schwab, C., High dimensional finite elements for elliptic problems with multiple scales. *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal* **3** (2005), 168–194.

[28] Houston, P., Rannacher, R., Süli, E, A posteriori error analysis for stabilised finite element approximations of transport problems. *Comput. Methods Appl. Mech. Engrg.* **190** (11–12) (2000), 1483–1508.

[29] Houston, P., Schwab, C., Süli, E., Stabilized $hp$-finite element methods for first-order hyperbolic problems. *SIAM J. Numer. Anal.* **37** (2000), 1618–1643.

[30] Houston, P., Schwab, C., Süli, E., Discontinuous $hp$-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.* **39** (2002), 2133–2163.

[31] Houston, P., Senior, B., Süli, E., $hp$-Discontinuous Galerkin finite element methods for hyperbolic problems: error analysis and adaptivity. *Int. J. Numer. Meth. Fluids* **40** (2002), 153–169.

[32] Houston, P., Senior, B., Süli, E., Sobolev regularity estimation for $hp$-adaptive finite element methods. In *ENUMATH 2001, European Conference on Numerical Mathematics and Applications* (F. Brezzi et al., eds.), Springer-Verlag, Berlin 2003, 631–656.

[33] Houston, P., Süli, E., Stabilized $hp$-finite element approximation of partial differential equations with nonnegative characteristic form. *Computing* **66** (2001), 99–119.

[34] Houston, P., Süli, E., $hp$-Adaptive discontinuous Galerkin finite element methods for hyperbolic problems. *SIAM J. Sci. Comp.* **23** (2001), 1225–1251.

[35] Hörmander, L., *The Analysis of Linear Partial Differential Operators II: Differential Operators with Constant Coefficients*. Reprint of the 1983 edition, Springer-Verlag, Berlin 2005.

[36] Hughes, T. J. R., Brooks, A. N., A multidimensional upwind scheme with no crosswind diffusion. In *Finite Element Methods for Convection Dominated Flows* (T. J. R. Hughes, ed.), AMD 34, ASME, New York 1979.

[37] Johnson, C., Nävert, U., Pitkäranta, J., Finite element methods for linear hyperbolic problems. *Comp. Meth. Appl. Mech. Engrg.* **45** (1984), 285–312.

[38] Johnson, C., Saranen, J., Streamline diffusion method for the incompressible Euler and Navier–Stokes equations. *Math. Comp.* **47** (1986), 1–18.

[39] Johnson, C., Schatz, A., Wahlbin, L., Crosswind smear and pointwise errors in the streamline diffusion finite element methods. *Math. Comp.* **49** (1987), 25–38.

[40] van Kampen, N. G., *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam 1992.

[41] Lapeyre, B., Pardoux, É., Sentis, R., *Introduction to Monte-Carlo Methods for Transport and Diffusion Equations*. Oxford Texts in Applied and Engineering Mathematics, Oxford University Press, Oxford 2003.

[42] Larson, M. G., Barth, T. J., A posteriori error estimation for discontinuous Galerkin approximations of hyperbolic systems. In *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. Karniadakis, and C.-W. Shu, eds.), Lecture Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin 2000.

[43] Laurençot, P., Mischler, S., The continuous coagulation fragmentation equations with diffusion. *Arch. Rational Mech. Anal.* **162** (2002) 45–99.

[44] Le Bris, C., Lions, P.-L., Renormalized solutions of some transport equations with $W^{1,1}$ velocities and applications. *Ann. Mat. Pura Appl.* (4) **183** (2004), 97–130.

[45] Nitsche, J.,Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg* **36** (1971), 9–15.

[46] Novak, E., Ritter, K., The curse of dimension and a universal method for numerical integration. In *Multivariate Approximation and Splines* (G. Nürnberger, J. Schmidt and G. Walz, eds.), International Series in Numerical Mathematics, Birkhäuser, Basel 1998, 177–188.

[47] Oden, J. T., Babuška, I., Baumann, C., A discontinuous *hp*-FEM for diffusion problems. J. Comput. Phys. **146** (1998), 491–519.

[48] Oleinik, O. A., Radkevič, E. V., *Second Order Equations with Nonnegative Characteristic Form*. Amer. Math. Soc., Providence, R.I., 1973.

[49] Öttinger, H.-C., *Stochastic Processes in Polymeric Fluids*. Springer-Verlag, Berlin 1996.

[50] von Petersdorff, T., Schwab, C., Numerical solution of parabolic equations in high dimensions. *M2AN Math. Model. Numer. Anal.* **38** (2004), 93–128.

[51] Prudhomme, S., Pascal, F., Oden, J. T., Romkes, A., Review of *a priori* error estimation for discontinuous Galerkin methods. TICAM Report 00–27, Texas Institute for Computational and Applied Mathematics, 2000.

[52] Reed, W. H., Hill, T. R., Triangular Mesh Methods for the Neutron Transport Equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[53] Rivière, B., Wheeler, M. F., A posteriori error estimates and mesh adaptation strategy for discontinuous Galerkin methods applied to diffusion problems. TICAM Report 00–10, Texas Institute for Computational and Applied Mathematics, 2000.

[54] Roos, H.-G., Stynes, M., Tobiska, L., *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*. Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin 1996.

[55] Schwab, C., *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics.* Numerical Methods and Scientific Computation, Clarendon Press, Oxford, 1998.

[56] Smolyak, S., Quadrature and interpolation formulas for products of certain classes of functions. *Soviet Math. Dokl.* **4** (1963), 240–243; Russian original in *Dokl. Akad. Nauk SSSR* **148** (1963), 1042–1045.

[57] Süli, E., Finite element approximation of high-dimensional transport-dominated diffusion problems. Oxford University Computing Laboratory, Numerical Analysis Technical Report Series, 05/19. In *Foundations of Computational Mathematics, Santander 2005* (L.-M. Pardo, A. Pinkus, E. Süli and M. Todd, eds.), London Math. Soc. Lecture Note Ser. 331, Cambridge University Press, Cambridge 2006, 343–370.

[58] Süli, E., Houston, P., Adaptive Finite Element Approximation of Hyperbolic Problems. In *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics* (T. Barth and H. Deconinck, eds.), Lecture Notes Comput. Sci. Eng. 25, Springer-Verlag, Berlin 2002, 269–344.

[59] Süli, E., Houston, P., Schwab, C., *hp*-Finite element methods for hyperbolic problems. In *The Mathematics of Finite Elements and Applications X. MAFELAP 1999* (J. R. Whiteman, ed.), Elsevier, Oxford 2000, 143–162.

[60] Süli, E., Schwab, C., Houston, P., *hp*-DGFEM for Partial Differential Equations with Nonnegative Characteristic Form. In *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. Karniadakis, and C.-W. Shu, eds.), Lecture Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin 2000, 221–230.

[61] Temlyakov, V., Approximation of functions with bounded mixed derivative. *Proc. Steklov Inst. Math.* **178** (1989)

[62] Wasilkowski, G., Woźniakowski, H., Explicit cost bounds of algorithms for multivariate tensor product problems. *J. Complexity* **11** (1995), 1–56.

[63] Zenger, C., Sparse grids. In *Parallel Algorithms for Partial Differential Equations.* (W. Hackbusch, ed.), Notes Numer. Fluid Mech. 31, Vieweg, Braunschweig, Wiesbaden 1991, 241–251.

University of Oxford, Computing Laboratory, Wolfson Building, Parks Road,
Oxford OX1 3QD, United Kingdom

E-mail: endre.suli@comlab.ox.ac.uk

# Ergodic control of diffusion processes

Vivek S. Borkar*

**Abstract.** Results concerning existence and characterization of optimal controls for ergodic control of nondegenerate diffusion processes are described. Extensions to the general 'controlled martingale problem' are indicated, which cover in particular degenerate diffusions and some infinite dimensional problems. In conclusion, some related problems and open issues are discussed.

**Mathematics Subject Classification (2000).** Primary 93E20; Secondary 60H20.

**Keywords.** Controlled diffusions, ergodic control, stationary Markov control, controlled martingale problems, dynamic programming.

## 1. Introduction

Ergodic or 'long run average' control of Markov processes considers the minimization of a time-averaged cost over admissible controls. This stands apart from the usual 'integral' cost criteria such as finite horizon or infinite horizon discounted cost criteria because neither the dynamic programming principle nor the usual 'tightness' arguments for existence of optima common to these set-ups carry over easily to the ergodic problem. Thus entirely new proof techniques have to be employed. The situation gets more complicated for continuous time continuous state space processes, of which diffusion processes are a prime example, because of the additional technicalities involved. This article describes first the reasonably well-understood case of non-degenerate diffusions, and then the partly resolved case of the more general 'controlled martingale problem' which covers degenerate diffusions and partially observed diffusions, among others.

An extended account of this topic will appear in [2].

## 2. Ergodic control of non-degenerate diffusions

**2.1. Preliminaries.** The $d$-dimensional ($d \geq 1$) controlled diffusion process $X(\cdot) = [X_1(\cdot), \ldots, X_d(\cdot)]^T$ is described by the stochastic differential equation

$$X(t) = X_0 + \int_0^t m(X(s), u(s))\, ds + \int_0^t \sigma(X(s))\, dW(s), \tag{1}$$

*The work is supported in part by a grant from the Department of Science and Technology, Government of India

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

for $t \geq 0$. Here:

1. for a compact metric 'control space' $U$, $m(\cdot,\cdot) = [m_1(\cdot,\cdot), \ldots, m_d(\cdot,\cdot)]^T : \mathcal{R}^d \times U \to \mathcal{R}^d$ is continuous and Lipschitz in the first argument uniformly with respect to the second,
2. $\sigma(\cdot) = [[\sigma_{ij}(\cdot)]]_{1 \leq i,j \leq d} : \mathcal{R}^d \to \mathcal{R}^{d \times d}$ is Lipschitz,
3. $X_0$ is an $\mathcal{R}^d$-valued random variable with a prescribed law $\pi_0$,
4. $W(\cdot) = [W_1(\cdot), \ldots, W_d(\cdot)]^T$ is a $d$-dimensional standard Brownian motion independent of $X_0$,
5. $u(\cdot) : \mathcal{R}^+ \to U$ is the 'control process' with measurable paths, satisfying the *non-anticipativity condition*: for $t > s \geq 0$, $W(t) - W(s)$ is independent of $\{X_0, W(y), u(y), y \leq s\}$. (In other words, $u(\cdot)$ does not anticipate the future increments of $W(\cdot)$.)

This class of $u(\cdot)$ is referred to as *admissible* controls. It is known that without loss of generality, one may take these to be adapted to the natural filtration of $X(\cdot)$, given by $\mathcal{F}_t$ = the completion of $\bigcap_{s>t} \sigma(X(y), y \leq s)$. We shall say that it is a *stationary Markov* control if in addition $u(t) = v(X(t)), t \geq 0$, for a measurable $v : \mathcal{R}^d \to U$. By a standard abuse of terminology, we identify this control with the map $v(\cdot)$. We shall say that (1) is *non-degenerate* if the least eigenvalue of $\sigma(\cdot)\sigma^T(\cdot)$ is uniformly bounded away from zero, *degenerate* otherwise. We use the 'weak solution' framework, i.e., only the law of the pair $(X(\cdot), u(\cdot))$ is prescribed and 'uniqueness' is interpreted as uniqueness in law. For this section, we assume non-degeneracy. This in particular implies existence of a unique *strong* solution for stationary Markov controls.

We shall also need the relaxation of the notion of control process $u(\cdot)$ above to that of a *relaxed* control process. That is, we assume that $U = \mathcal{P}(U_0)$, the space of probability measures on $U_0$ with Prohorov topology, where $U_0$ is compact metrizable (whence so is $U$) and $m_i(\cdot,\cdot)$, $1 \leq i \leq d$, are of the form

$$m_i(x,u) = \int \bar{m}_i(x,y)u(dy), \quad 1 \leq i \leq d,$$

for some $\bar{m}_i : \mathcal{R}^d \times U_0 \to \mathcal{R}$ that are continuous and Lipschitz in the first argument uniformly w.r.t. the second. We may write $u(t) = u(t, dy)$ to underscore the fact that it is a measure-valued process. Likewise for stationary Markov controls, write $v(\cdot) = v(\cdot, dy)$. Then the original notion of $U_0$-valued control $u_0(\cdot)$ (say) corresponds to $u(t, dy) = \delta_{u_0(t)}(dy)$, the Dirac measure at $u_0(t)$, for all $t$. We call such controls as *precise* controls. Precise stationary Markov controls may be defined accordingly.

The objective of ergodic control is to minimize

$$\limsup_{T \to \infty} \frac{1}{T} \int_0^T E[k(X(t), u(t))]\, dt \tag{2}$$

(the average version), or to a.s. minimize

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T k(X(t), u(t))\,dt \tag{3}$$

(the 'almost sure' version). Here $k : \mathcal{R}^d \times U \to \mathcal{R}$ is continuous. In view of our relaxed control framework, we take it to be of the form $k(x,u) = \int \bar{k}(x,y)u(dy)$ for a continuous $\bar{k} : \mathcal{R}^d \times U_0 \to \mathcal{R}$. This cost criterion is popular in applications where transients are fast, hence negligible, and one is choosing essentially from among the attainable 'steady states'. As mentioned above, we consider the non-degenerate case first. Most of the results presented in the remainder of this section have been established for bounded coefficients in the original sources, but the extension to the Lipschitz coefficients (implying linear growth) is not difficult and appears in [2]. One usually assumes (and we do) that $k$ is bounded from below.

**2.2. Existence results.** Let $v(\cdot)$ be a stationary Markov control such that the corresponding $X(\cdot)$ is positive recurrent and therefore has a unique stationary distribution $\eta^v \in \mathcal{P}(\mathcal{R}^d)$. Define the corresponding ergodic occupation measure as $\mu^v(dx,dy) = \eta^v(dx)v(x,dy)$. Costs (2), (3) will then equal ('a.s.' in the latter case) $\int \bar{k}d\mu^v$. A key result is:

**Theorem 2.1** ([18]). *The set $\mathcal{G} = \{\mu^v : v(\cdot)$ stationary Markov$\}$ is closed convex in total variation norm topology, with its extreme points corresponding to precise stationary Markov controls.*

We can say much more: define the empirical measures $\{\nu_t\}$ by:

$$\int f\,d\nu_t = \frac{1}{t}\int_0^t \int f(X(s), y)u(s,dy)\,ds, \quad f \in C_b(\mathcal{R}^d \times U_0),\ t > 0.$$

Let $\overline{\mathcal{R}} = \mathcal{R}^d \cup \{\infty\}$ = the one point compactification of $\mathcal{R}^d$ and view $\nu_t$ as a random variable in $\mathcal{P}(\overline{\mathcal{R}} \times U_0)$ that assigns zero mass to $\{\infty\} \times U_0$.

**Theorem 2.2** ([16]). *As $t \to \infty$, almost surely*

$$\nu_t \to \{\nu : \nu(A) = a\nu'(A \cap (\{\infty\} \times U_0)) + (1-a)\nu''(A \cap (\mathcal{R}^d \times U_0)) \text{ for all } A \text{ Borel in } \overline{\mathcal{R}} \times U_0,\ \text{with } a \in [0,1],\ \nu' \in \mathcal{P}(\{\infty\} \times U_0),\ \nu'' \in \mathcal{G}\}.$$

There are two important special cases for which Theorem 2.1 allows us to reduce the control problem to the infinite dimensional linear programming problem of minimizing $\int \bar{k}d\mu$ over $\mathcal{G}$ and thereby deduce the existence of an optimal precise stationary Markov control for the 'a.s.' version of the ergodic control problem [16]:

1. under a suitable 'stability condition' (such as a convenient 'stochastic Liapunov condition') that ensures compactness of $\mathcal{G}$ and a.s. tightness of $\{\nu_t\}$, or,

2. under a condition that penalizes escape of probability mass to infinity, such as the 'near-monotonicity condition':

$$\liminf_{||x||\to\infty} \min_u k(x,u) > \beta,$$

where $\beta =$ the optimal cost.

The latter condition is often satisfied in practice. The 'average' version of the ergodic cost can be handled similarly, using the average empirical measures $\{\bar{\nu}_t\}$ defined via

$$\int f\, d\bar{\nu}_t = \frac{1}{t}\int_0^t E\left[\int f(X(s),y)u(s,dy)\right] ds, \quad f \in C_b(\mathcal{R}^d \times U_0),\ t > 0,$$

in place of $\{\nu_t\}$.

**2.3. Dynamic programming.** The standard approach to dynamic programming for ergodic control, inherited from earlier developments in the discrete time and state problems, is to treat it as a limiting case of the infinite horizon discounted cost problem as the discount vanishes. Hence we begin with the infinite horizon discounted cost

$$E\left[\int_0^\infty e^{-\alpha t} k(X(t),u(t))\, dt\right],$$

where $\alpha > 0$ is the discount factor. Define

$$Lf(x,u) = \langle \nabla f(x), m(x,u)\rangle + \frac{1}{2}\mathrm{tr}\big(\sigma(x)\sigma^T(x)\nabla^2 f(x)\big)$$

for $f \in C^2(\mathcal{R}^d)$. We may write $L_u f(x)$ for $Lf(u,x)$, treating $u$ as a parameter. The Hamilton–Jacobi–Bellman (HJB) equation for the 'value function'

$$V^\alpha(x) = \inf E\left[\int_0^\infty e^{-\alpha t} k(X(t),u(t))\, dt | X(0) = x\right]$$

(where the infimum is over all admissible controls) can be arrived at by standard dynamic programming heuristic and is

$$\min_u (k(x,u) - \alpha V^\alpha(x) + LV^\alpha(x,u)) = 0$$

on the whole space. For $k$ bounded from below, $V^\alpha$ is its least solution in $C^2(\mathcal{R}^d)$. Define $\overline{V}^\alpha = V^\alpha - V^\alpha(0)$. Then $\overline{V}^\alpha$ satisfies

$$\min_u (k(x,u) - \alpha \overline{V}^\alpha(x) - \alpha V^\alpha(0) + L\overline{V}^\alpha(x,u)) = 0. \tag{4}$$

Under suitable technical conditions (such as near-monotonicity or stability conditions mentioned above) one can show that as $\alpha \downarrow 0$, $\overline{V}^\alpha(\cdot)$ and $\alpha V^\alpha(0)$ converge along

a subsequence to some $V, \beta$ in an appropriate Sobolev space and $\mathcal{R}$, respectively. Letting $\alpha \downarrow 0$ along this subsequence in (4), these are seen to satisfy

$$\min_u (k(x,u) - \beta + LV(x,u)) = 0.$$

This is the HJB equation of ergodic control. One can show uniqueness of $\beta$ as being the optimal ergodic cost and of $V$ up to an additive scalar in an appropriate function class depending on the set of assumptions one is working with. A verification theorem holds, i.e., the optimal stationary Markov control $v(\cdot)$ is characterized by the condition

$$v(x) \in \text{Argmin}\,(k(x,\cdot) + \langle \nabla V(x), m(x,\cdot)\rangle)\,, \text{ a.e.}$$

See [6], [17]. Note that the minimum will be attained in particular at a precise stationary Markov control, establishing the existence of an optimal precise stationary Markov control.

One also has the following stochastic representations for the ergodic value function $V$ (modulo an additive constant):

$$V(x) = \lim_{r \downarrow 0} \left( \inf E\left[ \int_0^{\tau_r} (k(X(s), u(s)) - \beta)\, ds | X(0) = x \right] \right),$$

where $\tau_r = \min\{t > 0 : ||X(t)|| = r\}$ for $r > 0$ [17] and the infimum is over all admissible controls. Alternatively,

$$V(x) = \inf \left( \inf_\tau E\left[ \int_0^{\tau} (k(X(s), u(s)) - \beta)\, ds | X(0) = x \right] \right),$$

where the inner infimum is over all bounded stopping times w.r.t. the natural filtration $\{\mathcal{F}_t\}$ of $X(\cdot)$, and the outer infimum is over all $\{\mathcal{F}_t\}$-adapted controls [21].

## 3. Controlled martingale problems

**3.1. Preliminaries.** Such explicit results are not as forthcoming in the more general scenario we discuss next. We shall denote by $E$ the Polish space that will serve as the state space of the controlled Markov process $X(\cdot)$, and by $U_0$ the compact metric 'control' space. $\mathcal{U}$ will denote the space of measurable maps $[0, \infty) \to U = \mathcal{P}(U_0)$ with the coarsest topology that renders continuous each of the maps

$$\mu(\cdot) = \mu(\cdot, du) \in \mathcal{U} \mapsto \int_0^T g(t) \int_{U_0} h(u) \mu(t, du)\, dt,$$

for all $T > 0$, $g \in L_2[0, T]$, $h \in C_b(U_0)$. This is compact metrizable (see, e.g., [9]). The control process $u(\cdot)$ can then be viewed as a $\mathcal{U}$-valued random variable.

For $\{f_k\}$, $f \in B(E) \stackrel{\text{def}}{=}$ the space of bounded measurable maps $E \to \mathcal{R}$, say that $f_k \xrightarrow{\text{bp}} f$ (where 'bp' stands for '*bounded pointwise*') if $\sup_{x,k} |f_k(x)| < \infty$ and

$f_k(x) \to f(x)$ for all $x$. $Q \subset B(E)$ is *bp-closed* if $f_k \in Q$ for all $k$ and $f_k \xrightarrow{\text{bp}} f$ together imply $f \in Q$. For $Q \subset B(E)$, define bp-closure$(Q)$ = the smallest bp-closed subset of $B(E)$ containing $Q$.

Let $A$ be an operator with domain $\mathcal{D}(A) \subset C_b(E)$ and range $\mathcal{R}(A) \subset C_b(E \times U_0)$. Let $\nu \in \mathcal{P}(E)$.

**Definition 3.1.** *An $E \times U$-valued process $(X(\cdot), \pi(\cdot) = \pi(\cdot, du))$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is said to be a solution to the relaxed controlled martingale problem for $(A, \nu)$ with respect to a filtration $\{\mathcal{F}_t, t \geq 0\}$ if:*

- *$(X(\cdot), \pi(\cdot))$ is $\{\mathcal{F}_t\}$-progressive;*
- *$\mathcal{L}(X(0)) = \nu$;*
- *for $f \in \mathcal{D}(A)$,*

$$f(X(t)) - \int_0^t \int_{U_0} Af(X(s), u)\pi(s, du)\, ds, \quad t \geq 0, \tag{5}$$

*is an $\{\mathcal{F}_t\}$-martingale.*

We omit explicit mention of $\{\mathcal{F}_t\}$ or $\nu$ when they are apparent from the context. The operator $A$ is assumed to satisfy the following conditions:

1. (C1) There exists a countable subset $\{g_k\} \subset \mathcal{D}(A)$ such that

$$\{(g, Ag) : g \in \mathcal{D}(A)\} \subset \text{bp-closure}(\{(g_k, Ag_k) : k \geq 1\}).$$

2. (C2) $\mathcal{D}(A)$ is an algebra that separates points in $E$ and contains constant functions. Also, $A\mathbf{1} = 0$, where $\mathbf{1}$ is the constant function identically equal to 1.
3. (C3) For each $u \in U_0$, let $A^u f(\cdot) = Af(\cdot, u)$. Then there exists an r.c.l.l. solution to the martingale problem for $(A^u, \delta_x)$ for all $u \in U_0$, $x \in E$.

For example, the following can be shown to fit this framework:

1. $X(\cdot)$ as in (1) with or without the non-degeneracy condition.
2. An important instance of the above is the 'separated control problem' for control of diffusions with partial observations, which we describe in some detail next. Append to (1) the 'observation equation'

$$Y(t) = \int_0^t h(X(s))\, ds + W'(t),$$

where $h : \mathcal{R}^d \to \mathcal{R}^s$ $(s \geq 1)$ is a Lipschitz observation map and $W'(\cdot)$ is an $s$-dimensional standard Brownian motion independent of $(X_0, W(\cdot))$, representing the (integrated) observation noise. The control $u(\cdot)$ is ideally

required to be adapted to the natural filtration of $Y(\cdot)$, but a standard relaxation allows for somewhat more general 'wide sense admissible' controls. These require merely that under a locally (in time) absolutely continuous change of measure that retains (1) but renders $Y(\cdot)$ itself an $s$-dimensional standard Brownian motion independent of $(X_0, W(\cdot))$, the future increments $Y(t+\cdot) - Y(t)$ should be independent of $\{X_0, W(\cdot), u(s), Y(s), s \le t\}$ for all $t > 0$. The correct state variable for this problem (to be precise, one choice thereof) turns out to be the $\mathcal{P}(\mathcal{R}^d)$-valued process $\{\mu_t\}$ of regular conditional laws of $X(t)$ given $\{Y(s), u(s), s \le t\}$ for $t \ge 0$. This evolves according to the equations of nonlinear filtering:

$$\mu_t(f) = \mu_0(f) + \int_0^t \mu_s(L_{u(s)}f)\,ds + \int_0^t \langle \mu_s(fh) - \mu_s(f)\mu_s(h), d\hat{Y}(s)\rangle \tag{6}$$

for $f \in C_b^2(\mathcal{R}^d)$, where we follow the notation $\nu(f) = \int f d\nu$. The products in the integrand of the stochastic integral in (6) are componentwise, and the process $\hat{Y}(t) = Y(t) - \int_0^t \mu_s(h)ds$, $t \ge 0$, is the so called 'innovations process' which is an $s$-dimensional standard Brownian motion that generates the same natural filtration as $Y(\cdot)$ [1]. The well-posedness of (6) can be established under additional regularity conditions on $h$ [24]. In terms of $\{\mu_t\}$, the ergodic cost can be rewritten as

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t E[\mu_s(\bar{k}(\cdot, u(s)))]\,ds.$$

The $\mathcal{P}(\mathcal{R}^d)$-valued controlled Markov process $\{\mu_t\}$ with this cost functional can be shown to fit the above framework. This is called the 'separated control problem' because it separates the issues of estimation and control.

3. Certain Hilbert-space valued controlled stochastic evolution equations can also be shown to fit the above framework [7].

**3.2. The control problem.** Let $k : E \times U_0 \to [0, \infty]$ be a continuous *running cost* function. The ergodic control problem is to minimize the *ergodic cost*

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t E\left[\int_{U_0} k(X(s), u)\pi(s, du)\right] ds. \tag{7}$$

We assume that the set of laws of $(X(\cdot), \pi(\cdot))$ for which this is finite is nonempty.

For a stationary $(X(\cdot), \pi(\cdot))$, define the associated *ergodic occupation measure* $\varphi \in \mathcal{P}(E \times U_0)$ by:

$$\int f(x, u)\varphi(dxdu) = E\left[\int_{U_0} f(X(t), u)\pi(t, du)\right].$$

Note that (7) then becomes $\int k d\varphi$. Let $\mathcal{G}$ denote the set of all ergodic occupation measures. From [7], we then have (see [27], [33], [34] for related results):

**Theorem 3.2.** *$\mathcal{G}$ is closed convex and is characterized as*

$$\mathcal{G} = \left\{\mu \in \mathcal{P}(E \times U_0) : \int Af \, d\mu = 0 \text{ for all } f \in \mathcal{D}(A)\right\}.$$

*In particular, for each $\mu \in \mathcal{G}$, there exists a stationary pair $(X(\cdot), \pi(\cdot))$ whose marginal at each time is $\mu$. Furthermore, $\pi(\cdot)$ may be taken to be stationary Markov.*

This can be made a starting point for existence results in specific cases. For example, for degenerate diffusions and the separated control problem for partially observed diffusions, somewhat stronger variants of the 'stability' and 'near-monotonicity' conditions described earlier suffice for the existence of an optimal stationary pair $(X(\cdot), \pi(\cdot))$. By considering the ergodic decomposition thereof, 'stationary' here may be improved to 'ergodic' [7]. Also, in view of the above theorem, the control therein may be taken to be stationary Markov.

This, however, does not imply that the process $X(\cdot)$ itself is time-homogeneous Markov, or even Markov. To establish the existence of an optimal Markov solution, we assume the following:

*For a fixed initial law $\nu$ of $X_0$, the attainable laws of $(X(\cdot), \pi(\cdot))$ form a tight set*

$$\bar{\mathcal{M}}(\nu) \subset \mathcal{P}(D([0, \infty); E) \times \mathcal{U}).$$

Simple sufficient conditions for this can be given in specific cases mentioned above. An immediate consequence of this is that $\bar{\mathcal{M}}(\nu)$ is in fact a compact convex set. Consider the equivalence relation on $\bar{\mathcal{M}}(\nu)$ that equates two laws when the corresponding one dimensional marginals agree for a.e. $t$. The set of equivalence classes, called the 'marginal classes', then forms a convex compact set in the quotient topology.

**Theorem 3.3.** *Every representative of an extremal marginal class corresponds to a Markov process.*

This is proved for degenerate diffusions in [10] and for the separated control problem in [20], but the same arguments carry over to the general case. This can be combined with the above to deduce the existence of an optimal pair $(X(\cdot), \pi(\cdot))$ such that $\pi(\cdot)$ is stationary Markov and $X(\cdot)$ Markov, though not necessarily time-homogeneous Markov [7]. Also, $(X(\cdot), \pi(\cdot))$ need not be stationary. Our experience with the non-degenerate case, however, suggests the existence of a stationary ergodic time-homogeneous Markov solution that is optimal. Under additional technical conditions, such a result has been proved in [8] by stretching the 'vanishing discount' argument, but there is scope for improvement.

As for dynamic programming, scattered results are available in specific cases. The degenerate problem has been approached in the viscosity solution framework [3], [4], [5]. For the separated control problem under partial observations, a martingale

dynamic programming principle has been derived [12], [13]. Dualizing the linear programme above yields the following dual linear programme that can be interpreted as 'dynamic programming inequalities' [7]:

*Maximize* $z \in \mathcal{R}$ *subject to* $Lf(x,u) + k(x,u) \geq z$, *for all* $x \in E$, $u \in U_0$, $f \in \mathcal{D}(L)$.

## 4. Some related problems and open issues

1. 'Ergodic control with constraints' seeks to minimize one ergodic cost functional while imposing bounds on one or more additional ergodic cost functionals. In the linear programming formulation alluded to above, this amounts to a few additional constraints. Existence of optimal precise stationary Markov controls has been proved in the non-degenerate case under suitable stability or near-monotonicity hypotheses [11], [18]. A Lagrange multiplier formulation can be used to aggregate the costs into a single cost.

2. We did not include control in the diffusion matrix $\sigma(\cdot)$. The reason for this is that, for stationary Markov controls $u(\cdot) = v(X(\cdot))$, one is in general obliged to consider at best measurable $v(\cdot)$. For a merely measurable diffusion matrix, even in the non-degenerate case only the existence of a weak solution is available, the uniqueness may not hold [26] (except in one and two dimensions – see [35], pp. 192–194). It may, however, be possible to work with 'the set of all weak solutions' in place of '*the*' solution, but this is not very appealing unless one has a good selection criterion that prescribes a unique choice from among the many.

3. Singularly perturbed ergodic control concerns ergodic control of diffusions wherein some components move on a much faster time scale, characterized by a perturbation parameter $\epsilon > 0$. One can show that as $\epsilon \downarrow 0$, the slower components satisfy an 'averaged' dynamics wherein the coefficients in their dynamics are averaged over the stationary distribution of the fast components when the latter is derived by 'freezing' the slower components to constant values. The ergodic control problem for this limiting case is then a valid approximation for the original problem for small $\epsilon > 0$. See [15] for a precise statement and proofs.

4. We have not considered several related problems with a similar flavor, such as ergodic control of reflected [14] or switching diffusions [23], [30], ergodic impulse control [32], singular ergodic control [31], and stochastic differential games with ergodic payoffs [19]. The latter in particular are also of interest in risk-sensitive control problems on infinite time horizon, which effectively get converted to two person zero sum stochastic differential games with ergodic payoffs after the celebrated 'log-transform' of the value function [22].

5. We have also not addressed the computational issues here. Two major strands therein are Markov chain approximations [28] and approximations of the infinite dimensional linear programmes [25].

## References


[1] Allinger, D. F., Mitter, S. K., New results on the innovations problem of nonlinear filtering. *Stochastics* **4** (1981), 339–348.

[2] Arapostathis, A., Borkar, V. S., Ghosh, M. K., *Ergodic Control of Diffusion Processes*. Book in preparation.

[3] Arisawa, M., Ergodic problem for the Hamilton-Jacobi-Bellman equation I. Existence of the ergodic attractor. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **14** (1997), 415–438.

[4] Arisawa, M., Ergodic problem for the Hamilton-Jacobi-Bellman equation II. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **15** (1998), 1–24.

[5] Basak, G., Borkar, V. S., Ghosh, M. K., Ergodic control of degenerate diffusions. *Stochastic Anal. Appl.* **15** (1997), 1–17.

[6] Bensoussan, A., Frehse, J., On Bellman equations of ergodic control in $R^n$. *J. Reine Angew. Math.* **429** (1992), 125–160.

[7] Bhatt, A. G., Borkar, V. S., Occupation measures for controlled Markov processes: characterization and optimality. *Ann. Probab.* **24** (1996), 1531–1562.

[8] Bhatt, A. G., Borkar, V. S., Existence of optimal Markov solutions for ergodic control of Markov processes. *Sankhya* **67** (2005), 1–18.

[9] Borkar, V. S., *Optimal Control of Diffusion Processes*. Pitman Research Notes in Math. 203, Longman Scientific and Technical, Harlow, UK, 1989.

[10] Borkar, V. S., On extremal solutions to stochastic control problems. *Appl. Math. Optim.* **24** (1991), 317–330.

[11] Borkar, V. S., Controlled diffusions with constraints II. *J. Math. Anal. Appl.* **176** (1993), 310–321.

[12] Borkar, V. S., Dynamic programmimg for ergodic control with partial observations. *Stochastic Process. Appl.* **103** (2003), 293–310.

[13] Borkar, V. S., Budhiraja, A., A further remark on dynamic programming for partially observed Markov processes. *Stochastic Process. Appl.* **112** (2004), 79–93.

[14] Borkar, V. S., Budhiraja, A., Ergodic control for constrained diffusions: characterization using HJB equations. *SIAM J. Control Optim.* **43** (2004), 1463–1492.

[15] Borkar, V. S., Gaitsgory, V., Singular perturbations in ergodic control of diffusions. Submitted.

[16] Borkar, V. S., Ghosh, M. K., Ergodic control of multidimensional diffusions I: the existence results. *SIAM J. Control Optim.* **26** (1988), 112–126.

[17] Borkar, V. S., Ghosh, M. K., Ergodic control of multidimensional diffusions II: adaptive control. *Appl. Math. Optim.* **21** (1990), 191–220.

[18] Borkar, V. S., Ghosh, M. K., Controlled diffusions with constraints. *J. Math. Anal. Appl.* **152** (1990), 88–108.

[19] Borkar, V. S., Ghosh, M. K., Stochastic differential games: an occupation measure based approach. *J. Optim. Theory Appl.* **73** (1992), 359–385; erratum *ibid.* **88**, 251–252.

[20] Borkar, V. S., Kumar, S., On extremal solutions of controlled non-linear filtering equations. *SIAM J. Control Optim.* **33** (1995), 718–724.

[21] Borkar, V. S., Mitter, S. K., A note on stochastic dissipativeness. In *Directions in Mathematical Systems Theory and Optimization* (ed. by A. Rantzer and C. I. Byrnes), Lecture Notes in Control and Inform. Sci. 286, Springer-Verlag, Berlin, Heidelberg 2003, 41–49.

[22] Fleming, W. H., McEneaney, W. M., Risk-sensitive control on an infinite horizon. *SIAM J. Control Optim.* **33** (1995), 1881–1915.

[23] Ghosh, M. K., Arapostathis, A., and Marcus, S. I., Optimal control of switching diffusions with applications to flexible manufacturing systems. *SIAM J. Control Optim.* **31** (1993), 1183–1204.

[24] Haussmann, U. G., L'Equation de Zakai et le Problème Séparé du Contrôle Optimal Stochastique. In *Seminaire de Probabilites XIX* (ed. by J. Azéma and M. Yor), Lecture Notes in Math. 1123, Springer-Verlag, Berlin, Heidelberg 1985, 37–62.

[25] Helmes, K., Stockbridge, R. H., Numerical comparison of controls and verification of optimality for stochastic control problems. *J. Optim. Theory Appl.* **106** (2000), 107–127.

[26] Krylov, N. V., *Controlled Diffusion Processes*. Applications of Mathematics 14, Springer-Verlag, New York, Heidelberg, Berlin 1980.

[27] Kurtz, T. G., Stockbridge, R., Existence of Markov controls and characterization of Markov controls. *SIAM J. Control Optim.* **36** (1998), 609–653; erratum *ibid.* **37** (1999), 1310–1311.

[28] Kushner, H. J., Dupuis, P., *Numerical Methods for Stochastic Control Problems in Continuous Time*. 2nd edition, Appl. Math. (N.Y.) 24, Springer-Verlag, New York 2001.

[29] Menaldi, J.-L., Robin, M., Ergodic control of reflected diffusions with jumps. *Appl. Math. Optim.* **35** (1997), 117–137.

[30] Menaldi, J.-L., Perthame, B., Robin, M., Ergodic problem for optimal stochastic switching. *J. Math. Anal. Appl.* **147** (1990), 512–530.

[31] Menaldi, J.-L., Robin, M., Taksar, M. I., Singular ergodic control for multidimensional Gaussian processes. *Math. Control Signals Systems* **5** (1992), 93–114.

[32] Robin, M., On some impulse control problems with long run average cost. *SIAM J. Control Optim.* **19** (1981), 333–358.

[33] Stockbridge, R. H., Time-average control of martingale problems: existence of a stationary solution. *Ann. Probab.* **18** (1990), 190–205.

[34] Stockbridge, R. H., Time-average control of martingale problems: a linear programming formulation. *Ann. Probab.* **18** (1990), 206–217.

[35] Stroock, D. W., Varadhan, S. R. S., *Multidimensional Diffusion Processes*. Grundlehren Math. Wiss. 233, Springer-Verlag, Berlin, New York 1979.

School of Technology and Computer Science, Tata Institute of Fundamental Research,
Homi Bhabha Road, Mumbai 400005, India.
E-mail: borkar@tifr.res.in

# Convex optimization of graph Laplacian eigenvalues

Stephen Boyd*

**Abstract.** We consider the problem of choosing the edge weights of an undirected graph so as to maximize or minimize some function of the eigenvalues of the associated Laplacian matrix, subject to some constraints on the weights, such as nonnegativity, or a given total value. In many interesting cases this problem is convex, *i.e.*, it involves minimizing a convex function (or maximizing a concave function) over a convex set. This allows us to give simple necessary and sufficient optimality conditions, derive interesting dual problems, find analytical solutions in some cases, and efficiently compute numerical solutions in all cases.

In this overview we briefly describe some more specific cases of this general problem, which have been addressed in a series of recent papers.

- *Fastest mixing Markov chain.* Find edge transition probabilities that give the fastest mixing (symmetric, discrete-time) Markov chain on the graph.
- *Fastest mixing Markov process.* Find the edge transition rates that give the fastest mixing (symmetric, continuous-time) Markov process on the graph.
- *Absolute algebraic connectivity.* Find edge weights that maximize the algebraic connectivity of the graph (*i.e.*, the smallest positive eigenvalue of its Laplacian matrix). The optimal value is called the *absolute algebraic connectivity* by Fiedler.
- *Minimum total effective resistance.* Find edge weights that minimize the total effective resistance of the graph. This is same as minimizing the average commute time from any node to any other, in the associated Markov chain.
- *Fastest linear averaging.* Find weights in a distributed averaging network that yield fastest convergence.
- *Least steady-state mean-square deviation.* Find weights in a distributed averaging network, driven by random noise, that minimizes the steady-state mean-square deviation of the node values.

**Mathematics Subject Classification (2000).** Primary 05C35; Secondary 90C25.

**Keywords.** Graph theory, Laplacian matrix, convex optimization, semidefinite programming, Markov chain, distributed averaging, effective resistance.

*Supported in part by the MARCO Focus Center for Circuit and System Solutions (C2S2, www.c2s2.org) under contract 2003-CT-888, by AFOSR grant AF F49620-01-1-0365, by NSF grant ECS-0423905, and by DARPA/MIT grant 5710001848. This paper reports work with co-authors Persi Diaconis, Arpita Ghosh, Seung-Jean Kim, Sanjay Lall, Pablo Parrilo, Amin Saberi. Jun Sun, and Lin Xiao.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

## 1. Introduction

Let $G = (V, E)$ be an undirected graph with $n = |V|$ nodes and $m = |E|$ edges, with weights $w_1, \ldots, w_m \in \mathbb{R}$ on the edges. Suppose edge $l$ connects vertices (or nodes) $i$ and $j$. We define $a_l \in \mathbb{R}^n$ as $(a_l)_i = 1$, $(a_l)_j = -1$, with other entries 0. The *weighted Laplacian* (matrix) is the $n \times n$ matrix defined as

$$L = \sum_{l=1}^{m} w_l a_l a_l^T = A \operatorname{diag}(w) A^T,$$

where $\operatorname{diag}(w) \in \mathbb{R}^{m \times m}$ is the diagonal matrix formed from $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$, and $A \in \mathbb{R}^{n \times m}$ is the *incidence matrix* of the graph, $A = [a_1 \ \cdots \ a_m]$.

We assume that the weights are such that $L$ is positive semidefinite, which we write as $L \succeq 0$. This is always the case when the weights are nonnegative. Since $L\mathbf{1} = 0$, where $\mathbf{1}$ is the vector with all components one, $L$ has smallest eigenvalue 0, corresponding to the eigenvector $\mathbf{1}$. We denote the eigenvalues of the Laplacian matrix $L$ as

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Let $\phi$ be a symmetric closed convex function defined on a convex subset of $\mathbb{R}^{n-1}$. Then

$$\psi(w) = \phi(\lambda_2, \ldots, \lambda_n) \tag{1}$$

is a convex function of $w$ [2, §5.2]. Thus, a symmetric convex function of the positive Laplacian eigenvalues yields a convex function of the edge weights. As a simple example, consider $\phi(u_1, \ldots, u_{n-1}) = \sum_{i=1}^{n-1} u_i$, *i.e.*, the sum. In this case we have

$$\psi(w) = \sum_{i=2}^{n} \lambda_i = \sum_{i=1}^{n} \lambda_i = \operatorname{Tr} L = 2\mathbf{1}^T w,$$

twice the sum of the edge weights, which is linear and therefore also convex. As another example, the function $\phi(u_1, \ldots, u_{n-1}) = \max_{i=1}^{n-1} u_i$ (which is convex and symmetric) yields the function $\psi(w) = \lambda_n$, the largest eigenvalue (or spectral radius) of the Laplacian matrix (and a convex function of the edge weights).

We consider optimization problems with the general form

$$\begin{array}{ll} \text{minimize} & \psi(w) \\ \text{subject to} & w \in \mathcal{W}, \end{array} \tag{2}$$

where $\mathcal{W}$ is a closed convex set, and the optimization variable here is $w \in \mathbb{R}^m$. The problem (2) is to choose edge weights on a graph, subject to some constraints, in order to minimize a convex function of the positive eigenvalues of the associated Laplacian matrix. We can also handle the case of maximizing a concave function $\phi$ of the positive Laplacian eigenvalues, by minimizing $-\psi$ over $w \in \mathcal{W}$.

The problem (2) is a *convex optimization problem*. Roughly speaking, this means that the analysis of the problem is fairly straightforward, and that the problem is easily solved numerically; see, *e.g.*, [6]. In the cases we will consider, the problem (2) can be formulated even more specifically as a *semidefinite program* (SDP), which has the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \sum_{i=1}^n x_i A_i \preceq B. \end{array} \tag{3}$$

Here $x \in \mathbb{R}^n$ is the variable, and the problem data are $c \in \mathbb{R}^n$ and the symmetric matrices $A_1, \ldots, A_n, B \in \mathbb{R}^{k\times k}$. The inequality symbol $\preceq$ between symmetric matrices refers to inequality with respect to the cone of positive semidefinite matrices. The constraint $\sum_{i=1}^n x_i A_i \preceq B$ is called a *linear matrix inequality* (LMI). The SDP (3) can be thought of as a generalization of a linear program (LP),

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \sum_{i=1}^n x_i a_i \leq b, \end{array}$$

where here, $a_1, \ldots, a_n, b$ are vectors, and the inequality symbol between vectors means componentwise. Many results for LPs have analogs for SDPs; moreover, in the last 15 years or so, effective algorithms for numerically solving SDPs have been developed, and are now widely used in many application areas.

## 2. Fastest mixing Markov chain

In this section we briefly describe the problem of finding the fastest mixing symmetric Markov chain on a given graph. Many more details (and additional references) can be found in [4, 5].

We consider a symmetric Markov chain on the graph $G$, with transition matrix $P \in \mathbb{R}^{n\times n}$, where $P_{ij} = P_{ji}$ is the probability of a transition from vertex $i$ to vertex $j$. Since $P$ is symmetric, the uniform distribution $\pi_{\text{unif}} = (1/n)\mathbf{1}^T$ is an equilibrium distribution. The rate of convergence of the distribution $\pi(t)$ to uniform is governed by $\mu(P)$, the *second largest eigenvalue magnitude* (SLEM) of $P$, with smaller $\mu(P)$ meaning faster asymptotic convergence. To find the fastest mixing symmetric Markov chain on the graph, we must choose the transition matrix $P$ to minimize $\mu(P)$, subject to the following conditions:

$$P = P^T, \quad P\mathbf{1} = \mathbf{1}, \quad P_{ij} \geq 0, \;\; i, j = 1, \ldots, n, \quad P_{ij} = 0 \;\; \text{for } (i, j) \notin E.$$

The first three conditions state that $P$ is a symmetric stochastic matrix; the last states that transitions can only occur over the graph edges.

Identifying the graph edge weights with edge transition probabilities, we find that $P$ can be expressed as $P = I - L$. The conditions above are equivalent to the conditions

$$w \geq 0, \quad \text{diag}(L) \leq \mathbf{1}$$

imposed on the edge weight vector $w$. (Here $\mathrm{diag}(L)$ is the vector consisting of the diagonal entries of $L$, and both inequalities above are vector inequalities, *i.e.*, componentwise.)

The eigenvalues of $P$ are $1-\lambda_1, \ldots, 1-\lambda_n$. Since $1-\lambda_1 = 1$, and $|1-\lambda_i| \leq 1$ (since $P$ is stochastic), its SLEM is given by

$$\mu(P) = \max\{|1-\lambda_2|, \ldots, |1-\lambda_n|\} = \max\{1-\lambda_2, \lambda_n - 1\}. \tag{4}$$

This has the general form (1), with $\phi(u_1, \ldots, u_{n-1}) = \max_{i=1}^{n-1} |1-u_i|$. In particular, the SLEM $\mu(P)$ is a convex function of the edge transition probabilities. Thus, the fastest mixing symmetric Markov chain problem can be expressed as our general problem (2), with $\mathcal{W} = \{w \mid w \geq 0,\ \mathrm{diag}(L) \leq \mathbf{1}\}$, a polyhedron.

The semidefinite programming formulation of the problem is

$$\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & -\gamma I \preceq I - L - (1/n)\mathbf{1}\mathbf{1}^T \preceq \gamma I, \quad w \geq 0, \quad \mathrm{diag}(L) \leq \mathbf{1}, \end{array}$$

with variables $w \in \mathbb{R}^m$ and $\gamma \in \mathbb{R}$.

Since the fastest mixing symmetric Markov chain problem is convex, indeed, equivalent to an SDP, it can be solved effectively. Generic methods can be used for problems with only a few thousand edges; far larger problems, with millions of edges, can be solved using subgradient optimization techniques, exploiting Lanczos methods to efficiently compute a few extreme eigenvalues and eigenvectors of $I - L - (1/n)\mathbf{1}\mathbf{1}^T$; see [5].

The optimal transition probabilities can be quite interesting; for example, a graph can have many edges with optimal transition probability zero. This means (roughly) that those edges are not needed to achieve fastest mixing on the given graph. We also note that the optimal transition probabilities can yield a mixing rate that is unboundedly better than some simple standard schemes for assigning transition probabilities for fast mixing, such as the maximum-degree method, or the Metropolis–Hastings method [5].

Standard methods can be used to construct various dual problems for the fastest mixing symmetric Markov chain problem. One such dual is

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T z \\ \text{subject to} & Y\mathbf{1} = 0, \quad Y = Y^T, \quad \|Y\|_* \leq 1 \\ & (z_i + z_j)/2 \leq Y_{ij}, \quad (i,j) \in E, \end{array} \tag{5}$$

with variables $z \in \mathbb{R}^n$ and $Y \in \mathbb{R}^{n\times n}$. Here $\|Y\|_* = \sum_{i=1}^n |\lambda_i(Y)|$, the sum of the singular values of $Y$, which is the dual norm of the spectral norm. This dual problem is convex, since the objective, which is maximized, is linear, hence concave, and the constraints are all convex. We have the following:

- *Weak duality.* If $Y$, $z$ are feasible for the dual problem (5), then we have $\mathbf{1}^T z \leq \mu^\star$, where $\mu^\star$ is the optimal value of the fastest mixing symmetric Markov chain problem.

- *Strong duality.* There exist $Y^\star$, $z^\star$ that are optimal for the dual problem, and satisfy $\mathbf{1}^T z^\star = \mu^\star$. This means that the optimal values of the primal and dual problems are the same, and that the dual problem yields a sharp lower bound on the optimal SLEM.

Both of these conclusions follow from general results for convex optimization problems (see, *e.g.*, [10, 1, 6]). We can conclude strong duality using (a refined form of) Slater's condition (see, *e.g.*, [1, §3.3] and [6, §5.2]), since the constraints are all linear equalities and inequalities.

## 3. Fastest mixing Markov process

Here we briefly describe the problem of finding the fastest mixing continuous-time symmetric Markov process on a given graph [11].

Consider a continuous-time Markov process on the graph $G$, with transition rate (or intensity) $w_l$ across edge $l$. The probability density $\pi(t) \in \mathbb{R}^{1\times n}$ at time $t \geq 0$ is given by $\pi(t) = \pi(0)e^{-tL}$. It follows that the asymptotic rate of convergence to the uniform distribution is governed by $\lambda_2$, the smallest positive eigenvalue of the Laplacian matrix. The deviation from uniform distribution decays, in the worst case, as $e^{-\lambda_2 t}$. We can express $\lambda_2$ as

$$\lambda_2 = \min\{\lambda_2, \ldots, \lambda_n\},$$

which has the standard form (1), with $\phi(u_1, \ldots, u_{n-1}) = \min_{i=1}^{n-1} u_i$. Since the minimum function is concave, we see that $\lambda_2$ is a concave function of the edge weights $w$. It is evidently homogeneous in $w$, so to get a sensible problem we must normalize the weights in some way, for example, as $\mathbf{1}^T w = 1$.

To find the transition rates that give fastest convergence (among weights that sum to one), we pose the problem

$$\begin{array}{ll} \text{maximize} & \lambda_2 \\ \text{subject to} & w \geq 0, \quad \mathbf{1}^T w = 1, \end{array}$$

with variable $w \in \mathbb{R}^m$. This is a convex optimization problem, which can be formulated as the SDP

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & \gamma I \preceq L + \beta \mathbf{1}\mathbf{1}^T, \quad w \geq 0, \quad \mathbf{1}^T w = 1, \end{array}$$

with variables $\gamma, \beta \in \mathbb{R}$, $w \in \mathbb{R}^m$.

The same problem, allocating a fixed total edge weight across the graph edges so as to maximize the smallest positive Laplacian eigenvalue, arises in other areas. For example, $\lambda_2$ arises in graph theory, and is called the *algebraic connectivity* of the graph. Fiedler refers to the maximum value of $\lambda_2$ that can be obtained by allocating a fixed total weight to the edges of a graph, as its *absolute algebraic connectivity* [7].

The dual of the fastest mixing Markov process problem can be given a very interesting interpretation. It is equivalent to the following problem. We are given some distances $d_1, \ldots, d_m$ on the graph edges. The goal is find a configuration of points $x_1, \ldots, x_n \in \mathbb{R}^n$ that satisfy $\|x_i - x_j\|_2 \leq d_l$, whenever edge $l$ connects vertices $i$ and $j$, and in addition maximizes the total variance, given by $\sum_{i \neq j} \|x_i - x_j\|^2$. This problem was recently formulated in the machine learning literature as a method for identifying low dimensional structure in data; see, *e.g.*, [12].

## 4. Minimum total effective resistance

Here we describe the problem of choosing the edge weights to minimize the total effective resistance of a graph, subject to some given total weight [8]. We consider the graph as an electrical circuit or network, with the edge weight representing the conductance (inverse of resistance) of the associated electrical branch. We define $R_{ij}$ as the resistance in the network seen between nodes $i$ and $j$. The total effective resistance is defined as $R = \sum_{i<j} R_{ij}$.

The total effective resistance comes up in several applications beyond circuit theory. For example, it is proportional to the average commute time, over all pairs of vertices, in the random walk on the graph defined by the weights $w_l$ [8]. (The probability of a transition from vertex $i$ to vertex $j$ is $w_l$, the associated edge weight, divided by the total weight of all edges adjacent to vertex $i$.)

It can be shown that

$$R = \frac{1}{n} \sum_{i=2}^{n} 1/\lambda_i,$$

*i.e.*, it is proportional to the sum of the inverses of the positive Laplacian eigenvalues. This follows our general form (1), with $\phi(u_1, \ldots, u_{n-1}) = \sum_{i=1}^{n-1} 1/u_i$, with domain $\mathbb{R}^{n-1}_{++}$. ($\mathbb{R}_{++}$ is the set of positive reals.) In particular, the total effective resistance is a convex function of the weight vector $w$. Minimizing total effective resistance, subject to $w \geq 0$ and $\mathbf{1}^T w = 1$, is thus a convex optimization problem.

The problem can be formulated as the SDP

$$\begin{array}{ll} \text{minimize} & n \operatorname{Tr} Y \\ \text{subject to} & \mathbf{1}^T w = 1, \quad w \geq 0, \\ & \begin{bmatrix} L + (1/n)\mathbf{1}\mathbf{1}^T & I \\ I & Y \end{bmatrix} \succeq 0, \end{array}$$

with variables $w \in \mathbb{R}^m$ and the (slack) matrix $Y = Y^T \in \mathbb{R}^{n \times n}$ (see [8]).

## 5. Fast averaging

Here we describe the problem of choosing edge weights that give fastest averaging, using a classical linear iteration [13]. The nodes start with value $x(0) \in \mathbb{R}^n$, and at each iteration we update the node values as $x(t+1) = (I - L)x(t)$. The goal is to choose the edge weights so that $x_i(t)$ converges, as rapidly as possible, to the average value, *i.e.*, $x(t) \to (1/n)\mathbf{1}\mathbf{1}^T x(0)$.

This iteration can be given a very simple interpretation. At each step, we replace each node value with a weighted average of its previous value and its neighbors' previous values. The weights used to form the average are taken from the graph edge weights, with the self-weight chosen so that the sum of the adjacent edge weights, plus the self-weight, equals one. The weights used to carry out this local averaging sum to one at each node, but can be negative.

When the weights are symmetric (which we assume here), the convergence rate of this averaging process is determined by the SLEM of $I - L$, *i.e.*, (4), exactly as in the Markov chain problem. The difference here is that the weights can be negative; in the Markov chain, of course, the weights (transition probabilities) must be nonnegative. The optimal weights can be found by solving the unconstrained problem

$$\text{minimize} \quad \max_{i=2}^{n} |1 - \lambda_i|,$$

which evidently is a convex optimization problem. It can be posed as the SDP

$$\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & -\gamma I \preceq L - (1/n)\mathbf{1}\mathbf{1}^T \preceq \gamma I, \end{array}$$

with variables $\gamma \in \mathbb{R}$, $w \in \mathbb{R}^m$. Without loss of generality, we can assume that $L \succeq 0$. The problem is the same as the fastest mixing symmetric Markov chain problem, but without the nonnegativity requirement on $w$. It often happens that some of the optimal weights are negative [13].

## 6. Minimum RMS consensus error

Here we describe a variation on the fastest linear averaging problem described above, in which an additive random noise perturbs the node values [15]. The iteration is $x(t+1) = (I - L)x(t) + v(t)$, where $v(t) \in \mathbb{R}^n$ are uncorrelated zero mean unit variance random variables, *i.e.*,

$$\mathbf{E}\, v(t) = 0, \quad \mathbf{E}\, v(t)v(t)^T = I, \quad \mathbf{E}\, v(t)v(s)^T = 0, \quad t \neq s.$$

This iteration arises in noisy averaging, distributed data fusion, and load balancing applications; see the references in [15].

We can measure the effectiveness of the averaging iteration at countering the effects of the additive noises by the steady-state mean-square deviation, defined as

$$\delta_{\mathrm{ss}} = \lim_{t\to\infty} \mathbf{E}\left(\frac{1}{n}\sum_{i<j}(x_i(t) - x_j(t))^2\right).$$

The steady-state mean-square deviation can be expressed as

$$\delta_{\mathrm{ss}} = \sum_{i=2}^{n} \frac{1}{\lambda_i(2-\lambda_i)},$$

provided $0 < \lambda_i < 2$ for $i = 2, \ldots, n$, and is infinite otherwise. (The condition $0 < \lambda_i < 2$ for $i = 2, \ldots, n$ is the same as $\max\{1-\lambda_2, \lambda_n - 1\} < 1$, which is the condition that the linear iteration for averaging, without the additive noise, converges.) Once again, this has the standard form (1), with $\phi(u_1, \ldots, u_{n-1}) = \sum_{i=1}^{n-1} 1/(u_i(2-u_i))$, with domain $(0,2)^{n-1}$. In particular, we see that $\delta_{\mathrm{ss}}$ is a convex function of the edge weights. To find the weights that yield the smallest steady-state mean-square deviation, we simply minimize the convex function $\delta_{\mathrm{ss}}$ over $w \in \mathbb{R}^m$.

## 7. Methods

All the problems described above can be effectively solved numerically, by a variety of standard methods for convex optimization, including interior-point methods for modest sized problems (with a few thousand weights) and subgradient-based methods for larger problems. We can exploit structure in the problems (such as sparsity of the underlying graph) to increase the efficiency of these methods.

We can also exploit symmetry in solving the problems. Two edges are symmetric if there exists an automorphism of the graph that maps one edge to the other. Whenever two edges are symmetric, we can assume without loss of generality that the corresponding edge weights are equal. (This follows from a basic result in convex optimization: there is always a solution that is invariant under the group of permutations that leave the objective function and constraint set fixed.) If the symmetry group of the graph is large, this can considerably reduce the size of the optimization problem that needs to be solved. As an extreme example, consider an edge-transitive graph, *i.e.*, one in which any two edges are symmetric. For such a graph, we can assume that all edge weights are equal, *i.e.*, there is only one common edge weight to be determined. This reduces the problem to one with at most one scalar variable (the common edge weight); if there is an equality constraint, such as $\mathbf{1}^T w = 1$, we conclude that an optimal solution is given by uniform edge weights, $w = (1/m)\mathbf{1}$ [3]. This idea is used in [9] to reduce some specific weight optimization problems to ones with a handful of variables, which can be solved analytically.

## References


[1] Bertsekas, D. P., *Nonlinear Programming*. Second edition, Athena Scientific, Belmont, MA, 1999.

[2] Borwein, J., and Lewis, A., *Convex Analysis and Nonlinear Optimization, Theory and Examples*. CMS Books Math./Ouvrages Math. SMC 3, Springer-Verlag, New York 2000.

[3] Boyd, S., Diaconis, P., Parrilo, P., and Xiao, L., Symmetry analysis of reversible Markov chains. *Internet Math.* **2** (1) (2005), 31–71.

[4] Boyd, S., Diaconis, P., Sun, J., and Xiao, L., Fastest mixing Markov chain on a path. *Amer. Math. Monthly* **113** (1) (2006), 70–74.

[5] Boyd, S., Diaconis, P., and Xiao, L., Fastest mixing Markov chain on a graph. *SIAM Review* **46** (4) (2004), 667–689.

[6] Boyd, S., and Vandenberghe, L., *Convex Optimization*. Cambridge University Press, Cambridge 2004.

[7] Fiedler, M., Absolute algebraic connectivity of trees. *Linear and Multilinear Algebra* **26** (1990), 85–106.

[8] Ghosh, A., and Boyd, S., Minimizing effective resistance of a graph. *SIAM Review*, to appear; www.stanford.edu/~boyd/eff_res.

[9] Ghosh, A., and Boyd, S., Upper bounds on algebraic connectivity via convex optimization. *Linear Algebra Appl.*, to appear; www.stanford.edu/~boyd/eigbounds_laplacian.

[10] Rockafellar, R. T., *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[11] Sun, J., Boyd, S., Xiao, L., and Diaconis, P., The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, to appear 2006; www/~boyd/fmmp.

[12] Weinberger, K., and Saul, L., Unsupervised learning of image manifolds by semidefinite programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR-04*), Washington D.C., 2004.

[13] Xiao, L., and Boyd, S., Fast linear iterations for distributed averaging. *Systems Control Lett.* **53** (2004), 65–78.

[14] Xiao, L., and Boyd, S., Optimal scaling of a gradient method for distributed resource allocation. *J. Optim. Theory Appl.* **129** (3) (2006), to appear.

[15] Xiao, L., Boyd, S., and Kim, S.-J., Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, summitted May 2005; www/~boyd/lms_consensus.



Electrical Engineering Department, Stanford University, Packard 264, Stanford, CA 94305, U.S.A.

E-mail: boyd@stanford.edu

# Controllability of evolution equations of fluid dynamics

Oleg Yu. Emanouilov (Imanuvilov)

**Abstract.** In this paper we will discuss recent developments in controllability of evolution equations of fluid mechanics. The control is assumed to be distributed either on a part of the boundary or locally distributed in some subdomain. We will present some ideas of proof of main theorems. Special attention will be paid to the technique based on Carleman estimates.

**Mathematics Subject Classification (2000).** 93C20, 35B37.

**Keywords.** Carleman estimates, exact controllability, Navier–Stokes system.

## 1. Introduction

This paper is concerned with the problem of exact controllability of partial differential equations with control concentrated either on the part of the boundary or locally distributed inside of the boundary in some subdomain. The typical statement of general controllability problem, which we are going to discuss in this paper, can be formulated as follows: let a function $y(t, x)$, which describes the state of a system, satisfy a semilinear partial differential equation

$$\partial_t y + A(x, D)y + F(x, y, \nabla y) = \chi_\omega u \quad \text{in } (0, T) \times \Omega, \tag{1.1}$$

$$B(x, D)y = 0 \quad \text{on } (0, T) \times \partial\Omega, \quad y(0, \cdot) = y_0, \tag{1.2}$$

where $A(x, D)$ is a linear operator, $F$ is a nonlinear term, $B(x, D)$ is a boundary operator, $\chi_\omega$ is the characteristic function of the domain $\omega \subset \Omega$ where the control function $u(t, x)$ is supported.

The initial conditions, the function $y_0(x)$ and another function $y_1(x)$ called target function, are given. Let us choose some time moment $T$. Then the exact controllability problem may be formulated as follows: Find the control $u$ and the state function $y$ such that

$$y(T, \cdot) = y_1. \tag{1.3}$$

The solvability properties of the controllability problem (1.1)–(1.3) are completely different from the properties of the initial boundary value problem (1.1)–(1.2) with fixed $u$. For initial value problems with a reasonable choice of boundary conditions and for smooth initial conditions we usually expect the uniqueness of the solution. Moreover, if a priori estimates are obtained, this solution typically may be extended

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

globally in time. On the other hand, for most boundary/locally distributed controllability problems of equations of mathematical physics, solutions are not unique and a priori estimates typically are absent. In case of control of linear equations these difficulties do not produce a huge problem, since the controllability problem typically can be reduced to an *observability* problem which can be formulated as follows. Suppose two Banach spaces $X$ and $Y$ are given. For the solution of the adjoint linear equation

$$-\partial_t z + A^*(x, D)z = 0 \quad \text{in } (0, T) \times \Omega, \tag{1.4}$$

$$B^*(x, D)z = 0 \quad \text{on } (0, T) \times \partial\Omega \tag{1.5}$$

one needs to obtain the a priori estimate

$$\|z\|_X \le C\|\chi_\omega z\|_Y. \tag{1.6}$$

The initial conditions at $t = T$ for problem (1.4), (1.5) are not assumed to be known. This creates the main difficulty in proving estimate (1.6). There are several methods to deal with such observability problems:

1. The method based on the theorem on propagation of singularities (see Bardos–Lebeau–Rauch [2]);

2. Multipliers method (see [31], [28], [27], [29], [39]);

3. Carleman estimates (see [21], [22], [25], [26], [37], [38]).

The first two methods are effective for the wave and Schrödinger equations. As for the equations of parabolic type and the generalized Stokes system, the Carleman type estimates with the singular weight functions appears to be more effective method compared to methods 1 and 2.

## 2. Controllability of parabolic equations and the Burgers equation

In a bounded domain $\Omega \in \mathbb{R}^N$ with $\partial\Omega \in C^2$ we consider the semilinear parabolic equation

$$G(y) = \partial_t y - \Delta y + f(t, x, y) = \chi_\omega u + g \tag{2.7}$$

with given initial condition and zero Dirichlet boundary conditions

$$y|_{(0,T)\times\partial\Omega} = 0, \quad y(0, \cdot) = y_0. \tag{2.8}$$

Here $\omega$ is an arbitrary but fixed subdomain, $\chi_\omega$ is the characteristic function of the domain $\omega$ and $u(t, x)$ is the control locally distributed in $\omega$. Suppose that the target function $y_1(x)$ is given and some moment of time $T$ is fixed. We are looking for control $u$ such that

$$y(T, \cdot) = y_1. \tag{2.9}$$

Since the solution $y(t, x)$ of the heat equation with zero right-hand side is analytic as a function of $x$ for any positive $t$, we cannot solve in general problem (2.7)–(2.9) for an arbitrary smooth target function $y_1$.

Let us assume that

$$f \in C^1([0,T]\times\bar{\Omega}\times\mathbb{R}^1), \quad f(t,x,0)=0 \quad \text{for all } (t,x)\in(0,T)\times\Omega, \tag{2.10}$$

and that the function $f(t,x,y)$ satisfies the Lipschitz condition

$$|f(t,x,\zeta_1)-f(t,x,\zeta_2)|\le K|\zeta_1-\zeta_2| \quad \text{for all } (t,x)\in(0,T)\times\Omega,\ \zeta_1,\zeta_2\times\mathbb{R}^1, \tag{2.11}$$

where the constant $K$ is independent of $t, x, \zeta$.

We have

**Theorem 2.1** ([22]). *Let $y_1\equiv 0$ and conditions* (2.10), (2.11) *hold true. Suppose that there exists $\delta>0$ such that $e^{\frac{1}{(T-t)^{1+\delta}}}g\in L^2((0,T)\times\Omega)$. Then for any $y_0\in \mathring{W}_2^1(\Omega)$, there exists a solution $(y,u)\in W_2^{1,2}((0,T)\times\Omega)\times L^2((0,T)\times\omega)$ to problem* (2.7)–(2.9).

Here $W^{1,2}((0,T)\times\Omega)=\{y(t,x)\mid \partial_t y, \partial_x^\beta y\in L^2((0,T)\times\Omega) \text{ for all } |\beta|\le 2\}$.

Thanks to assumption (2.11) by standard methods of functional analysis the proof of Theorem 2.1 may be reduced to the question of solvability of the controllability problem for the linear parabolic equation

$$\begin{aligned} &\partial_t v-\Delta v+c(t,x)v=\chi_\omega\tilde{u}+\tilde{g} \quad \text{in } (0,T)\times\Omega,\\ &v|_{(0,T)\times\partial\Omega}=0, \quad v(0,\cdot)=v_0, \quad v(T,\cdot)=0, \end{aligned} \tag{2.12}$$

where $c\in L^\infty((0,T)\times\Omega)$. The solvability of problem (2.12) is equivalent to obtaining the observability estimate for the adjoint parabolic equation:

$$-\partial_t z-\Delta z+c(t,x)z=q \quad \text{in } [0,T]\times\Omega, \tag{2.13}$$

$$z|_{(0,T)\times\partial\Omega}=0. \tag{2.14}$$

The observability estimate for (2.13)–(2.14) can be proved using the technique of Carleman estimates. First we need to introduce some weight functions. Let $\psi\in C^2(\overline{\Omega})$ be such that

$$\psi(x)>0 \ \text{ for all } x\in\Omega, \quad \psi|_{\partial\Omega}=0, \quad |\nabla\psi(x)|>0 \ \text{ for all } x\in\Omega\setminus\omega_0, \tag{2.15}$$

where $\omega_0\subset\subset\omega$ is some open set. Using the function $\psi$ we construct three more functions: $\varphi(t,x)=e^{\lambda\psi(x)}/(t(T-t))$, $\alpha(t,x)=(e^{\lambda\psi}-e^{2\lambda||\psi||_{C(\overline{\Omega})}})/(t(T-t))$, and $\eta(t,x)=(e^{\lambda\psi}-e^{2\lambda||\psi||_{C(\overline{\Omega})}})/(\ell(t)(T-t))$ where $\ell\in C^\infty[0,T]$, $\ell(t)>0$ for any $t\in[0,T]$ and $\ell(t)=t$ for $t\in\left[\frac{3T}{4},T\right]$.

The following holds:

**Lemma 2.1** ([22]). *There exists a number $\hat{\lambda}>0$ such that for an arbitrary $\lambda\ge\hat{\lambda}$ there exists $s_0(\lambda)$ such that for each $s\ge s_0(\lambda)$ solutions to problem* (2.13)–(2.14)

*satisfy the following inequality:*

$$\int_{(0,T)\times\Omega}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial z}{\partial t}\right|^2+|\Delta z|^2\right)+s\varphi|\nabla z|^2+s^3\varphi^3z^2\right)e^{2s\alpha}dx\,dt \\ \leq C\left(\int_{(0,T)\times\Omega}|q|^2e^{2s\alpha}dxdt+\int_{[0,T]\times\omega}s^3\varphi^3z^2e^{2s\alpha}dxdt\right), \tag{2.16}$$

*where the constant $C$ is independent of $s$.*

This estimate, combined with the standard energy estimate for equation (2.13), implies that for any $v_0\in\overset{\circ}{W}{}_2^1(\Omega)$ and $e^{-s_0\eta}\tilde{g}\in L^2((0,T)\times\Omega)$ there exists a solution to problem (2.12): a pair $(y,u)\in W^{1,2}((0,T)\times\Omega)\times L^2((0,T)\times\omega)$ such that $e^{-s_0\eta}\tilde{u}\in L^2((0,T)\times\Omega)$, $e^{-s_0\eta}v/(T-t)^{\frac{3}{2}}\in L^2((0,T)\times\Omega)$.

A different approach, still based on Carleman estimates, was proposed by G. Lebeau and L. Robbiano in [30] for linear parabolic equations with time independent coefficients. In [8], [11], [35] solutions for the controllability problem of the linear heat equation were constructed directly by solving a moment problem. In [36] the solution to the controllability problem for the heat equation was obtained from a solution of the corresponding problem for the wave equation. In [34] another method was proposed, essentially based on the solvability of the Cauchy problem for the one dimensional heat equation. Later this method was applied to the semilinear parabolic equation in [33]. The approximate controllability for equation (2.7) was proved in [9].

Next we consider the problem of exact controllability of equation (2.7) with boundary control. Let $\Gamma_0$ be an arbitrary subdomain of $\partial\Omega$. Suppose that the control $u$ is distributed over $\Gamma_0$:

$$G(y)=g,\quad y|_{(0,T)\times\Gamma_0}=u,\quad y|_{(0,T)\times\partial\Omega\setminus\Gamma_0}=0,\quad y(0,\cdot)=y_0,\quad y(T,\cdot)=y_1. \tag{2.17}$$

We have

**Theorem 2.2** ([22]). *Let $y_1\equiv 0$ and conditions* (2.10), (2.11) *hold true. Suppose that there exists $\delta>0$ such that $e^{\frac{1}{(T-t)^{1+\delta}}}g\in L^2((0,T)\times\Omega)$. Then for any $y_0\in\overset{\circ}{W}{}_2^1(\Omega)$ there exists a solution $(y,u)\in W_2^{1,2}((0,T)\times\Omega)\times L^2(0,T;H^{\frac{1}{2}}(\Gamma_0))$ to problem* (2.17).

Theorem 2.2 will easily follow from Theorem 2.1 if we enlarge the domain $\Omega$ up to $\tilde{\Omega}$ in such a way that

$$\omega\subset\tilde{\Omega},\quad \omega=\tilde{\Omega}\setminus\Omega,\quad \partial\omega\cup\partial\Omega\subset\Gamma_0.$$

Then we consider problem (2.7)–(2.9) in $\tilde{\Omega}$ with the control locally distributed in $\omega$. Since the existence of the solution $y$ is guaranteed by Theorem 2.1 we consider the restriction of $y$ on $\Omega$ and put $u=y|_{\Gamma_0}$.

Next we consider the situation when the target function is not zero. In order to solve the controllability problem we need some conditions on the functions $y_1$ and $g$.

**Condition 2.1.** There exists a constant $\tau > 0$ and a function $\tilde{u} \in L^2((0,T)\times\omega)$ such that the boundary value problem

$$G(\tilde{y}) = \chi_\omega \tilde{u} + g \quad \text{in } [T-\tau, T]\times\Omega, \quad \tilde{y}|_{[T-\tau,T]\times\partial\Omega} = 0, \quad \tilde{y}(T,\cdot) = y_1$$

has a solution $\tilde{y} \in W^{1,2}((0,T)\times\Omega)$.

We have

**Theorem 2.3** ([22]). *Let $y_0 \in \mathring{W}_2^1(\Omega)$ and $g \in L^2((0,T)\times\Omega)$. Suppose that* (2.10), (2.11) *hold true. Let the functions $y_1$ and $g$ satisfy Condition* 2.1. *Then there exists a solution $(y,u) \in W^{1,2}((0,T)\times\Omega)\times L^2((0,T)\times\omega)$ of problem* (2.7)–(2.9).

Theorem 2.3 provides necessary and sufficient conditions for solvability of problem (2.7)–(2.9).

A similar result holds true for the situation when the control is locally distributed over the boundary.

**Condition 2.2.** There exists a constant $\tau > 0$ and a function $\tilde{u} \in L^2((0,T)\times\omega)$ such that the boundary value problem

$$\begin{aligned} &G(\tilde{y}) = g \quad \text{in } [T-\tau, T]\times\Omega, \\ &\tilde{y}|_{[T-\tau,T]\times\Gamma_0} = \tilde{u}, \quad \tilde{y}|_{[T-\tau,T]\times\partial\Omega\setminus\Gamma_0} = 0, \quad \tilde{y}(T,\cdot) = y_1 \end{aligned}$$

has a solution $\tilde{y} \in W^{1,2}((0,T)\times\Omega)$.

The following holds:

**Theorem 2.4** ([22]). *Let $y_0 \in \mathring{W}_2^1(\Omega)$ and $g \in L^2((0,T)\times\Omega)$. Suppose that* (2.10), (2.11) *hold true. Let the functions $y_1$ and $g$ satisfy Condition* 2.2. *Then there exists a solution $(y,u) \in W^{1,2}((0,T)\times\Omega)\times L^2(0,T; H^{\frac{1}{2}}(\partial\Omega))$ of problem* (2.17).

In case when the nonlinear term of the parabolic equation is superlinear the situation is different. For example, there exists $y_0 \in C^\infty(\bar{\Omega})$ and a time moment $\hat{T}$ which depends on $\Omega$ only, such that any solution for the initial value problem

$$\partial_t y - \Delta y + y^2 = 0 \quad \text{in } \Omega, \quad y(0,\cdot) = y_0, \quad y|_{(0,T)\times\partial\Omega} = u$$

will blow up at some time $\tau(u) < \hat{T}$. Hence we even cannot prevent a blowup by the boundary control. The similar question for the nonlinearity $f(t,x,y) = -y^3$ is open. If the nonlinear term has the form $f(t,x,y) = y^3$ for any $y_0 \in W_2^1(\Omega)\cap L^6(\Omega)$ and sufficiently regular $u$ (which satisfies the compatibility condition) a solution to the initial value problem

$$\partial_t y - \Delta y + y^3 = 0 \quad \text{in } \Omega, \quad y(0,\cdot) = y_0, \quad y|_{(0,T)\times\partial\Omega} = u$$

exists and satisfies the a priori estimate

$$\frac{d}{dt}\int_\Omega \rho^7(x) y^2(t,x)dx + \frac{1}{8}\int_\Omega \rho^7(x) y^4(t,x)dx \le C,$$

where $\rho \in C^2(\bar{\Omega})$ is an arbitrary function such that $\rho(x) > 0$ for each $x \in \Omega$, $\rho|_{\partial\Omega} = 0$, $|\nabla\rho|_{\partial\Omega} \neq 0$ and the constant $C$ depends on $\rho$ only. This estimate immediately implies that for some open set of target functions $y_1(x)$ in $L^2(\Omega)$ there is no solution to problem (2.17).

Let us consider the Burgers equation

$$\partial_t y - \partial_x^2 y + \partial_x y^2 = \chi_\omega u(t,x), \quad (t,x) \in [0,T] \times [0,L], \tag{2.18}$$

with zero Dirichlet boundary conditions and the initial condition

$$y(t,0) = y(t,L) = 0, \quad y(0,\cdot) = y_0. \tag{2.19}$$

Here $\omega \subset [0,L]$ is an arbitrary but fixed open set. We are looking for a control $u$ such that

$$y(T,\cdot) = y_1 \tag{2.20}$$

The following holds:

**Theorem 2.5** ([16]). *Let $y_1 \in \overset{\circ}{W}{}_2^1(0,L)$ be a steady-state solution to the Burgers equation and $y_0 \in \overset{\circ}{W}{}_2^1(0,L)$. Then there exists a time moment $T(y_1)$ such that the controllability problem* (2.18)–(2.20) *has a solution $(y,u) \in W^{1,2}((0,T)\times[0,L]) \times L^2((0,T)\times[0,L])$.*

Suppose that $\omega$ satisfies the following condition:

$$\text{there exists } b > 0 \text{ such that } \omega \subset (b,L). \tag{2.21}$$

We have

**Lemma 2.2** ([16]). *Let $y(t,x)$ be a solution to problem* (2.18), (2.19). *Denote $y_+(t,x) = \max(y(t,x),0)$. Then for arbitrary $N > 5$ the following estimate holds true:*

$$\frac{d}{dt}\int_0^b (b-x)^N y_+^4(t,x)dx < \gamma(N) b^{N-5}. \tag{2.22}$$

*Here $\gamma(N) > 0$ is a constant depending on $N$ only.*

The immediate consequence of (2.22) is the existence of an open set of target functions which is unreachable by means of the locally distributed control satisfying (2.21) or by means of the boundary control concentrated at $x = L$.

If condition (2.21) fails, we of course do not have the a priori estimate (2.22). In terms of the boundary control this situation corresponds to the case when the control is located at both endpoints of the segment $[0,L]$. By Hopf's transformation this problem might be reduced to the controllability problem of the one-dimensional heat equation with control located at both endpoints of the segment $[0,L]$ but with one additional constraint: control functions are nonnegative. Then from results of [1] it follows that for some initial condition $y_0$ the set of all reachable functions is not dense in $L^2(0,L)$. Later we will see that the controllability properties of the Burgers equation and the Navier–Stokes system are completely different.

## 3. Local controllability of the Navier–Stokes system

In [32] J.-L. Lions conjectured that the Navier–Stokes system with boundary or locally distributed control is globally approximately controllable. This paper inspired intensive research in the area. In this section we discuss the local controllability results for the Navier–Stokes system and the Boussinesq system.

Let us consider the Navier–Stokes system defined on the bounded domain $\Omega \subset \mathbb{R}^N$ $(N = 2, 3)$ with boundary $\partial\Omega \in C^2$

$$\partial_t y(t,x) - \Delta y(t,x) + (y, \nabla)\, y + \nabla p = f + \chi_\omega u \quad \text{in } \Omega, \quad \operatorname{div} y = 0, \tag{3.23}$$

$$y|_{(0,T)\times\partial\Omega} = 0, \quad y(0,\cdot) = y_0, \tag{3.24}$$

where $y(t,x) = (y_1(t,x), \dots, y_N(t,x))$ is the velocity of fluid, $p$ is the pressure. The density of external forces $f(t,x) = (f_1(t,x), \dots, f_N(t,x))$ and the initial velocity $y_0$ are given, $u(t,x)$ is a control distributed in some arbitrary but fixed subdomain $\omega$ of the domain $\Omega$.

Let $(\hat{y}(t,x), \hat{p}(t,x))$ be a solution of the Navier–Stokes equations with the right-hand side $f$ exactly the same as in (3.23):

$$\partial_t \hat{y} - \Delta\hat{y} + (\hat{y}, \nabla)\hat{y} + \nabla\hat{p} = f \quad \text{in } (0,T)\times\Omega, \quad \operatorname{div}\hat{y} = 0, \quad \hat{y}|_{(0,T)\times\partial\Omega} = 0 \tag{3.25}$$

close enough to the initial condition $y_0$ at the moment $t = 0$

$$\|y_0 - \hat{y}(0,\cdot)\|_V \le \varepsilon, \quad \text{(the parameter } \varepsilon \text{ is sufficiently small)} \tag{3.26}$$

where $V = \{y(x) = (y_1, \dots, y_N) \in (W_2^1(\Omega))^N : \operatorname{div} y = 0 \text{ in } \Omega,\ y|_{\partial\Omega} = 0\}$.

We are looking for a control $u$ such that, for a given $T > 0$, the following equality holds

$$y(T,\cdot) = \hat{y}(T,\cdot). \tag{3.27}$$

In order to formulate our results, we introduce the following functional spaces:

$$H = \{y(x) = (y_1, \dots, y_N) \in (L^2(\Omega))^N : \operatorname{div} y = 0,\ \text{in } \Omega,\ (y, \vec{n})|_{\partial\Omega} = 0\},$$

$$V^{1,2}((0,T)\times\Omega) = \{y(t,x) \in (W_2^{1,2}((0,T)\times\Omega))^N : \operatorname{div} y = 0 \text{ in } \Omega,\ y|_{\partial\Omega} = 0\},$$

where $\vec{n} = \vec{n}(x) = (n_1(x), \dots, n_N(x))$ is the outward unit normal to $\partial\Omega$.

Suppose that the function $\hat{y}$ has the following regularity properties:

$$\begin{aligned} &\hat{y} \in L^\infty((0,T)\times\Omega), \\ &\partial_t\hat{y} \in L^2(0,T; L^\sigma(\Omega)), \quad \sigma > 6/5 \text{ for } N = 3, \sigma > 1 \text{ for } N = 2. \end{aligned} \tag{3.28}$$

The following result in particular gives us a positive answer to the question of the possibility of stabilization of the flow near an unstable steady state solution by means of locally distributed control.

**Theorem 3.1** ([12]). *Let $y_0 \in V$, $f \in L^2(0,T;H)$ and suppose that the pair $(\hat{y},\hat{p})$ solves* (3.25) *and satisfies condition* (3.28). *Then for sufficiently small $\varepsilon > 0$ there exists a solution $(y,p,u) \in V^{1,2}((0,T)\times\Omega)\times L^2(0,T;W_2^1(\Omega))\times(L^2((0,T)\times\omega)))^N$ to problem* (3.23), (3.24), (3.26), (3.27).

This result first has been proved in [15] for the control distributed over the whole boundary $\partial\Omega$. In [21] the case of control distributed over an arbitrary small subdomain $\omega$, but with some assumptions on the geometry of $\Omega$ was considered. Finally, in [23], these assumptions on $\Omega$ were removed under the regularity condition on the function $\hat{y}$ which is stronger then (3.28).

Since the existence theorem 3.1 is local, in order to prove this existence result one first proves the solvability of the controllability problem for the Navier–Stokes equation linearized at trajectory $\hat{y}$:

$$\begin{cases} \partial_t\tilde{y} - \Delta\tilde{y} + (\hat{y},\nabla)\tilde{y} + (y,\nabla)\hat{y} + \nabla\tilde{p} = f + \chi_\omega\tilde{u},\ \operatorname{div}\tilde{y} = 0 & \text{in } (0,T)\times\Omega,\\ \tilde{y} = 0 & \text{on } (0,T)\times\partial\Omega,\\ \tilde{y}(0,\cdot) = y_0,\ \ \tilde{y}(T,\cdot) = 0 & \text{in } \Omega. \end{cases} \tag{3.29}$$

After the solvability of (3.29) is established in appropriate functional spaces the conclusion of the Theorem 3.1 follows from the standard implicit function theorem.

The typical way to solve (3.29) is to reduce it to the observability problem for the operator adjoint to the operator of the linearization of the Navier–Stokes system at trajectory $\hat{y}$. More precisely, let the function $z \in L^2(0,T;H)$ satisfy the equations

$$-\partial_t z - \Delta z - Dz\hat{y} = \nabla\pi + g \text{ in } (0,T)\times\Omega, \tag{3.30}$$

$$\operatorname{div} z = 0, \quad z|_{(0,T)\times\partial\Omega} = 0, \tag{3.31}$$

where the function $Dz = \nabla z + \nabla z^t$.

Denote $\alpha(t,x) = \frac{e^{\lambda\psi(x)+8\|\psi\|_{L^\infty(\Omega)}} - e^{10\lambda\|\psi\|_{L^\infty(\Omega)}}}{(t(T-t))^4}$, $\alpha^*(t) = \min_{x\in\Omega}\alpha(t,x)$, $\hat{\alpha}(t) = \max_{x\in\Omega}\alpha(t,x)$, $\hat{\varphi}(t,x) = \frac{e^{8\lambda\|\psi\|_{L^\infty(\Omega)}}}{(t(T-t))^4}$, $\varphi(t,x) = \frac{e^{8\lambda\|\psi\|_{L^\infty(\Omega)}+\psi(x)}}{(t(T-t))^4}$. The function $\psi$ is introduced in (2.15). For the system (3.30)–(3.31), we have the following observability estimate:

**Theorem 3.2** ([12]). *There exist three positive constants $\hat{s}$, $\hat{\lambda}$, $C$ depending on $\Omega$ and $\omega$ such that for every $z_0 \in H$, $g \in L^2((0,T)\times\Omega)$ the corresponding solution to* (3.30), (3.31) *verifies:*

$$\begin{aligned} &\int_{(0,T)\times\Omega}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial z}{\partial t}\right|^2 + \sum_{i,j=1}^n\left|\frac{\partial^2 z}{\partial x_i\partial x_j}\right|^2\right) + s\lambda^2\varphi|\nabla z|^2 + s^3\lambda^4\varphi^3|z|^2\right)e^{2s\alpha}dxdt\\ &\quad\le C(1+T^2)\left(s^{\frac{15}{2}}\lambda^{20}\int_{(0,T)\times\Omega}|g|^2\hat{\varphi}^{\frac{15}{2}}e^{4s\hat{\alpha}-2s\alpha^*}dxdt\right.\\ &\qquad\left.+\int_{(0,T)\times\omega}s^{16}\lambda^{40}\hat{\varphi}^{16}|z|^2e^{8s\hat{\alpha}-6s\alpha^*}dxdt\right) \quad \text{for all } s \ge s_0, \end{aligned} \tag{3.32}$$

*for all* $\lambda \geq \hat{\lambda}(1 + \|\hat{y}\|^2_{L^\infty((0,T)\times\Omega)} + \|\partial_t \hat{y}\|^2_{L^2(0,T;L^\sigma(\Omega))} + e^{\hat{\lambda}T\|\hat{y}\|^2_{L^\infty((0,T)\times\Omega)}})$ *and* $s \geq \hat{s}(T^4 + T^8)$.

The strategy of the proof of (3.32) is as follows. First we apply the Carleman estimate (2.16) to equations (3.31). Next we need to eliminate the norm of the function $\nabla\pi$ on the right-hand side. In order to do that we observe that the pressure $\pi$ for each $t \in [0, T]$ satisfies the Laplace equation

$$-\Delta\pi = \operatorname{div}(Dz\hat{y}) + \operatorname{div} g \quad \text{in } \Omega. \tag{3.33}$$

Since the velocity field $z$ satisfies the zero Dirichlet boundary conditions, there are no explicit boundary conditions for the pressure $\pi$. Therefore to equation (3.33) we apply the Carleman estimates for elliptic equations obtained in [24] with weights which minimize the contribution of the boundary terms. Finally we eliminate the norms of the functions $\pi|_{\partial\Omega}$ and $\chi_\omega\pi$ using some a priori estimates for the initial value problems for the Stokes system and the heat equation.

In many controllability problems in addition to be locally distributed in a subdomain, the control $u$ is required to satisfy some additional constraints. Below we discuss the situation when in problems (3.23), (3.24), (3.27) the control satisfies the following constraint: one of the components of the vector function $u(t, x)$ is identically equal zero on $(0, T) \times \Omega$. Suppose that $\omega$ satisfies the following condition:

$$\text{there exists } x^0 \in \partial\Omega,\ \tilde{\delta} > 0 \text{ such that } \overline{\omega} \cap \partial\Omega \supset B(x^0; \tilde{\delta}) \cap \partial\Omega. \tag{3.34}$$

($B(x^0; \tilde{\delta})$ is the ball centered at $x^0$ of radius $\tilde{\delta}$.)

Let $E = H$ for $N = 2$ and $E = H \cap L^4(\Omega)$ for $N = 3$. Assume that the initial condition $y_0$ is close to $\hat{y}(0, \cdot)$ in the norm of the space $E$:

$$\|y_0 - \hat{y}(0, \cdot)\|_E \leq \varepsilon. \tag{3.35}$$

We have

**Theorem 3.3.** *Assume that* $\omega$ *satisfies* (3.34). *Let* $y_0 \in E$, $f \equiv 0$ *and suppose that the pair* $(\hat{y}, \hat{p})$ *solves* (3.25) *and satisfies condition* (3.28). *Then for sufficiently small* $\varepsilon > 0$ *there exists a solution* $(y, p, u)$ *to problem* (3.23), (3.24), (3.35), (3.27) *with control* $u \in (L^2((0, T) \times \omega))^N$ *having one component identically zero.*

In the case of locally distributed control with zero component $u_k$ for the corresponding observability problem, associated with (3.30),(3.31) we do not have any information on the $k$-th component of the function $z$ in $(0, T) \times \omega$. This means that the function $z_k$ should not appear in the right-hand side of the inequality (3.32). This difficulty can be overcome if we recall that $z$ is divergence free function and therefore its $k$-th component satisfies the equation $\partial_{x_k} z_k = \sum_{j=1, j\neq k}^{N} \partial_{x_j} z_j$. From this ordinary differential equation, thanks to zero Dirichlet boundary conditions and assumption (3.34), in some subdomain of $\omega$ we can estimate $z_k$ by the remaining components of the function $z$ and then apply (3.32).

Next we consider the similar controllability problem of the Boussinesq system.

$$\begin{cases} \partial_t y - \Delta y + (y, \nabla) y + \nabla p = \chi_\omega u + \theta\, e_N, \ \ \operatorname{div} y = 0 & \text{in } (0,T)\times\Omega, \\ \partial_t \theta - \Delta\theta + (y, \nabla\theta) = \chi_\omega h & \text{in } (0,T)\times\Omega, \\ y = 0,\ \theta = 0 & \text{on } (0,T)\times\partial\Omega, \\ y(0,\cdot) = y_0,\ \theta(0,\cdot) = \theta_0 & \text{in } \Omega. \end{cases} \tag{3.36}$$

In the domain $(0,T)\times\omega$ we control both the density of external forces $u$ and the density of external heat sources $h$.

Let $(\hat y, \hat p, \hat\theta)$ be a sufficiently regular solution to the Boussinesq system:

$$\partial_t \hat y - \Delta \hat y + (\hat y, \nabla)\hat y + \nabla \hat p = \hat\theta\, e_N, \ \ \operatorname{div} \hat y = 0 \quad \text{in } (0,T)\times\Omega, \tag{3.37}$$

$$\partial_t \hat\theta - \Delta\hat\theta + (\hat y, \nabla\hat\theta) = 0 \quad \text{in } (0,T)\times\Omega, \tag{3.38}$$

$$\hat y = 0,\ \hat\theta = 0 \quad \text{on } (0,T)\times\partial\Omega, \tag{3.39}$$

$$\hat y(0,\cdot) = \hat y_0,\ \hat\theta(0,\cdot) = \hat\theta_0 \quad \text{in } \Omega. \tag{3.40}$$

Assume that $\hat y$ satisfies (3.28) and the temperature $\hat\theta$ has the following regularity

$$\begin{aligned} &\hat\theta \in L^\infty((0,T)\times\Omega), \\ &\partial_t\hat\theta \in L^2(0,T; L^\sigma(\Omega)), \quad \sigma > 1 \text{ if } N = 2, \sigma > 6/5 \text{ if } N = 3. \end{aligned} \tag{3.41}$$

In addition to condition (3.34) we assume that

$$\text{there exists } k < N, \text{ such that } n_k(x^0) \neq 0. \tag{3.42}$$

Our goal is to prove that for some $\varepsilon > 0$ , whenever $(y_0, \theta_0) \in E \times L^2(\Omega)$ and

$$\|(y_0,\theta_0) - (\hat y_0, \hat\theta_0)\|_{E\times L^2(\Omega)} \le \varepsilon, \tag{3.43}$$

we can find $L^2$ controls $u$ and $h$ with $u_k \equiv u_N \equiv 0$ such that

$$y(T,\cdot) = \hat y(T,\cdot) \text{ and } \theta(T,\cdot) = \hat\theta(T,\cdot) \text{ in } \Omega. \tag{3.44}$$

We note that for dimension $N = 2$ we are trying to control both the velocity field and the temperature by choosing the density of external heat sources in the subdomain $\omega$. The following holds:

**Theorem 3.4.** *Assume that $\omega$ satisfies* (3.34) *and* (3.42). *Let $y_0 \in E$, $\theta_0 \in L^2(\Omega)$ and suppose that the pair $(\hat y, \hat\theta, \hat p)$ solves* (3.37)–(3.40) *and satisfies conditions* (3.28), (3.41). *Then for sufficiently small $\varepsilon > 0$ there exists a solution $(y, \theta, p, u, h)$ to problem* (3.36), (3.43), (3.44) *such that $(u, h) \in (L^2((0,T)\times\Omega))^{N+1}$ and $u_k \equiv u_N \equiv 0$. In particular, if $N = 2$, we have local exact controllability with controls $u \equiv 0$ and $h \in L^2((0,T)\times\omega)$.*

## 4. Global controllability of the Navier–Stokes and the Boussinesq system

In this section we will discuss the global controllability of the Boussinesq and the Navier–Stokes systems. We start with the controllability problem for the Boussinesq system with periodic boundary conditions:

$$\partial_t y - \Delta y + (y, \nabla)y + \nabla p = f + \theta e_N + \chi_\omega u \quad \text{in } K = \Pi_{j=1}^N [0, 2\pi], \quad \operatorname{div} y = 0, \tag{4.45}$$

$$\partial_t \theta - \Delta\theta + (y, \nabla\theta) = g + \chi_\omega h \quad \text{in } K, \tag{4.46}$$

$$y(t, \dots x_i + 2\pi, \dots) = y(t, x), \quad \theta(t, \dots x_i + 2\pi, \dots) = \theta(t, x) \quad \text{for all } i \in \{1, \dots, N\}, \tag{4.47}$$

$$y(0, \cdot) = y_0, \quad \theta(0, \cdot) = \theta_0, \quad y(T, \cdot) = \hat{y}(T, \cdot), \quad \theta(T, \cdot) = \hat{\theta}(T, \cdot). \tag{4.48}$$

Here $\hat{\theta}$, $\hat{y}$ is some solution to the Boussinesq system with the same right-hand side:

$$\partial_t \hat{y} - \Delta\hat{y} + (\hat{y}, \nabla)\hat{y} + \nabla\hat{p} = f + \hat{\theta} e_N \quad \text{in } (0, T) \times K, \quad \operatorname{div} \hat{y} = 0, \tag{4.49}$$

$$\partial_t \hat{\theta} - \Delta\hat{\theta} + (\hat{y}, \nabla\hat{\theta}) = g \quad \text{in } (0, T) \times K, \tag{4.50}$$

$$\hat{y}(t, \dots x_i + 2\pi, \dots) = \hat{y}(t, x), \quad \hat{\theta}(t, \dots x_i + 2\pi, \dots) = \hat{\theta}(t, x) \quad \text{for all } i \in \{1, \dots, N\}. \tag{4.51}$$

A very essential role in controllability problems for the Navier–Stokes system and the Boussinesq system is played by the type of boundary conditions.

For the case of periodic boundary conditions the situation is understood much better than for the case of Dirichlet boundary conditions. One reason for this striking difference is that for the periodic case we can construct explicitly a set of nonzero solutions of the Boussinesq system

$$\partial_t \tilde{y} - \Delta\tilde{y} + (\tilde{y}, \nabla)\tilde{y} = \nabla\tilde{p} + \tilde{\theta} e_N + \chi_\omega \tilde{u} \quad \text{in } K, \ \operatorname{div} \tilde{y} = 0, \quad \tilde{y}(0, \cdot) = \tilde{y}(T, \cdot) = 0, \tag{4.52}$$

$$\tilde{y}(t, \dots x_i + 2\pi, \dots) = \tilde{y}(t, x), \quad \tilde{\theta}(t, \dots x_i + 2\pi, \dots) = \tilde{\theta}(t, x) \quad \text{for all } i \in \{1, \dots, N\}, \tag{4.53}$$

$$\partial_t \tilde{\theta} - \Delta\tilde{\theta} + (\tilde{y}, \nabla\tilde{\theta}) = 0 \quad \text{in } K, \quad \tilde{\theta}(0, \cdot) = \tilde{\theta}(T, \cdot) = 0 \tag{4.54}$$

in the form

$$\begin{aligned} &\tilde{y}(t, x) = m(t, x), \quad \tilde{\theta}(t, x) \equiv 0, \\ &m(t, \dots, x_i + 2\pi, \dots) = m(t, x) \quad \text{for all } i \in \{1, \dots, N\}, \end{aligned} \tag{4.55}$$

where $m(t, x) = \nabla\gamma(t, x)$ and $\Delta\gamma(t, \cdot) = 0$ in $K \setminus \omega$ for all $t \in [0, T]$ and $\gamma(0, \cdot) = \gamma(T, \cdot) = 0$. (Obviously for the Dirichlet boundary conditions the function $\gamma \equiv 0$ is the only possible choice!) Note that for any $\tilde{N}$ the functions $(\tilde{N}\tilde{y}, \tilde{N}\tilde{\theta}, \tilde{N}\tilde{p})$ also

solve (4.52)–(4.54) with some $\tilde{u}_{\tilde{N}}$. If we are looking for a solution of the problem (4.45), (4.46), (4.47), (4.48) in the form $(y, \theta) = (Y + \tilde{N}\tilde{m}, \theta)$ then in new equations for $(Y, \theta)$ the large parameter $\tilde{N}$ will appear. Therefore the next logical step in finding $(Y, \theta)$ is to solve a controllability problem associated to the transport equation. In order to do that we need to make a special choice of the vector field $m$. The following holds:

**Lemma 4.1** ([17]). *There exists a vector field* $m(t, x) = (m_1(t, x), \dots, m_N(t, x)) \in C^\infty([0, T] \times K)$ *such that*

$$\operatorname{div} m = 0 \text{ in } [0, T] \times K, \quad m(t, x) = \nabla\gamma(t, x) \quad \text{and} \quad \Delta\gamma = 0 \text{ in } [0, T] \times (K \setminus \omega),$$

*for arbitrary* $k \in \mathbb{N}$

$$m(0, x) \equiv m(T, x) \equiv 0, \quad \frac{\partial^k m(t, x)}{\partial t^k}|_{t=0} = \frac{\partial^k m(t, x)}{\partial t^k}|_{t=T} = 0,$$

*and the relation*

$$\{(t, x(t, x_0), t \in (0, T)\} \cap [0, T] \times \omega \neq \emptyset$$

*is valid for every* $x_0 \in K$, *where* $x(t, x_0)$ *is solution to the Cauchy problem*

$$\frac{d}{dt}x(t, x_0) = m(t, x(t, x_0)), \quad x(t, x_0)|_{t=0} = x_0.$$

*Moreover,* $x(T, x_0) = x_0$ *for each* $x \in K$. *Furthermore there exist a finite cover* $\{\mathcal{O}_j \ j = 1, \dots, J\}$ *of* $K$ *by open sets* $\mathcal{O}_j$ *and a number* $\hat{\delta} > 0$ *such that for each* $j$ *all the curves* $x(t, x_0)$, $x_0 \in \mathcal{O}_j$ *lie in* $\omega$ *for some time interval* $\hat{\delta}$.

In case we choose the vector field $m$ as in Lemma 4.1 the following controllability problem may be solved for all regular initial data $y_0, \theta_0$:

$$\partial_t r + (m, \nabla)r + (r, \nabla)m - \nabla q_1 = \chi_\omega \bar{u}, \ \operatorname{div} r = 0,$$

$$\partial_t z + (m, \nabla z) = \chi_\omega \bar{h},$$

$$r(t, \dots, x_i + 2\pi, \dots) = r(t, x), \ z(t, \dots, x_i + 2\pi, \dots) = z(t, x), \ \ i \in \{1, \dots, N\},$$

$$r(0, \cdot) = y_0, \quad z(0, \cdot) = z_0, \quad r(T, \cdot) = \hat{y}(\varepsilon T, \cdot), \quad z(T, \cdot) = \hat{\theta}(\varepsilon T, \cdot).$$

Finally one can construct an approximation for the solution to problem (4.45)–(4.48) in the form

$$y(t, x) = \frac{1}{\varepsilon} m\Big(\frac{t}{\varepsilon}, x\Big) + r\Big(\frac{t}{\varepsilon}, x\Big) + y_\varepsilon, \quad \theta(t, x) = z\Big(\frac{t}{\varepsilon}, x\Big) + \theta_\varepsilon, \tag{4.56}$$

$$u(t, x) = \frac{1}{\varepsilon}\bar{u}\Big(\frac{t}{\varepsilon}, x\Big) - \chi_\omega \Delta \frac{1}{\varepsilon} m\Big(\frac{t}{\varepsilon}, x\Big), \quad h = \frac{1}{\varepsilon}\bar{h}\Big(\frac{t}{\varepsilon}, x\Big). \tag{4.57}$$

Here the terms $y_\varepsilon$, $\theta_\varepsilon$ are small provided that $\varepsilon > 0$ is small. Of course, we do not have the exact equality $y(\varepsilon T, \cdot) = \hat{y}(\varepsilon T, \cdot)$ but the difference $y(\varepsilon T, \cdot) - \hat{y}(\varepsilon T, \cdot)$

can be made arbitrarily small.( This proves Lions global controllability conjecture for the Boussinesq and the Navier–Stokes system with control distributed on the whole boundary.) Then the local controllability result similar to Theorem 3.4 could be applied in order to switch to the trajectory $(\hat{y}, \hat{\theta})$.

The idea to construct a solution to the controllability problem in the form (4.56), (4.57) was proposed by J. M. Coron in [3], [4] for the two dimensional Navier–Stokes system and Euler equation. In particular, Coron proved that if the control acts on an arbitrary open subset of the boundary which meets any connected component of this boundary, then the 2-D Euler equations are exactly controllable. Later the proof was extended to 3-D Euler equation by Glass [18], [19]. In [6], Coron constructed explicitly the feedback laws which globally asymptotically stabilize the fluid flow described by the Euler equation. The global exact controllability of the Navier–Stokes system on a manifold without boundary was studied in [7].

In order to formulate controllability results for the Boussinesq system rigorously we introduce the functional spaces

$$V^0(K) = \{y(x) \in (L^2(K))^N : \operatorname{div} y = 0,\ y(t, \dots x_i + 2\pi, \dots) = y(t, x) \quad \text{for all } i \in \{1, \dots, N\}\},$$

$$V^1(K) = \{y \in V^0(K) \cap (W_2^1(K))^N\},$$

$$V^{1,2} = \{y(t, x) \in (W_2^{1,2}((0, T)\times K))^N : \operatorname{div} y = 0,\ y(t, \dots x_i + 2\pi, \dots) = y(t, x)\}.$$

We have the following result:

**Theorem 4.1** ([17]). *Let* $y_0 \in V^1(K)$, $\theta_0 \in W_2^1(K)$, $f \in L^2(0, T; V^0(K))$, $g \in L^2((0, T) \times K)$ *and suppose that for some* $\beta \in (0, 1)$ *the function* $(\hat{y}, \hat{\theta}, \hat{p}) \in C^1(0, T; V^0(K) \cap (C^{2,\beta}(K))^N) \times C^1(0, T; C^{2,\beta}(K)) \times L^2(0, T; W_2^1(K))$ *is a given solution of the Boussinesq system* (4.49)–(4.51). *Then there exists a solution* $(y, \theta, p, u, h) \in V^{1,2} \times W^{1,2}((0, T)\times K) \times L^2(0, T; W_2^1(K)) \times (L^2((0, T)\times\omega))^{N+1}$ *to problem* (4.45), (4.46), (4.47), (4.48).

Now we consider the problem of global controllability for the 2-D Navier–Stokes system with zero Dirichlet boundary conditions and control distributed over a part of the boundary. We need to introduce a "large parameter" in this problem but the analog of (4.55) for a general domain is hard to find. Therefore below we consider the Navier–Stokes system in the special domain $\Omega = \{(x_1, x_2) : x_1 \in (0, 1),\ x_2 \in (0, 1)\}$. Let us consider the following controllability problem:

$$\begin{cases} \partial_t y - \Delta y + (y, \nabla)y = \nabla p + f, \ \operatorname{div} y = 0 & (t, x) \in (0, T) \times \Omega, \\ y(t, 0, x_2) = 0 & (t, x_2) \in (0, T) \times (0, 1), \\ y(0, \cdot) = y_0, \ \ y(T, \cdot) = 0 & x = (x_1, x_2) \in \Omega. \end{cases} \tag{4.58}$$

The initial condition $y_0$ satisfies

$$\operatorname{div} y_0 = 0, \quad x \in \Omega, \qquad \text{and} \qquad y_0(0, x_2) = 0, \quad x_2 \in (0, 1). \tag{4.59}$$

Observe that in system (4.58) we did not fix traces of $y$ on $(\{1\} \times (0,1)) \cup ((0,1) \times \{0,1\})$. They can be chosen arbitrarily and considered as a boundary control.

Next we construct an analog of the vector field $m$. Let the function $U(t,x)$ have the form $U(t,x) = (0, z(t,x_1))$ where $z = z(t,x_1)$ solves the following problem associated to a linear heat equation:

$$\begin{cases} \partial_t z - \partial^2_{x_1 x_1} z = c(t) & (t,x_1) \in (0,T) \times (0,2), \\ z(t,0) = 0, \quad z(t,1) = w(t) & t \in (0,T), \\ z(0,x_1) = 0 & x_1 \in (0,2). \end{cases} \tag{4.60}$$

Here $c(t)$ is a constant for each $t$ such that

$$c(0) \neq 0, \quad w(t) \in C^\infty[0,T], \quad w(0) = 0,\ w'(0) = c(0),\ w''(0) = c'(0).$$

Using this function, we construct $U(t,x) = (0, z(t,x_1))$ and $q = x_2 c(t)$ for $t \in (0,T)$, $x \in \tilde{K} = [0,1] \times [0,2]$, which for an arbitrary $\tilde{N} \in \mathbb{R}^1$ solves

$$\begin{cases} \partial_t(\tilde{N}U) - \Delta \tilde{N}U + (\tilde{N}U, \nabla)(\tilde{N}U) = \nabla(\tilde{N}q), \\ \operatorname{div}(\tilde{N}U) = 0 & (t,x) \in (0,T) \times \tilde{K}, \\ \tilde{N}U(t,0,x_2) = 0 & (t,x_2) \in (0,T) \times \mathbb{R}^1, \\ \tilde{N}U(0,x) = 0 & x \in \tilde{K}. \end{cases} \tag{4.61}$$

We have

**Theorem 4.2** ([20]). *Let $f \in L^2((0,T) \times \Omega)$ and let $y_0 \in W^1_2(\Omega)$ satisfy* (4.59). *Then there exists a sequence of functions $f_\varepsilon$ such that*

$$f_\varepsilon \to f \ \text{ in } \ L^{p_0}(0,T;V'), \quad p_0 \in (1, 8/7),$$

*and there exists at least one solution to the controllability problem*

$$\begin{cases} \partial_t y_\varepsilon - \Delta y_\varepsilon + (y_\varepsilon, \nabla) y_\varepsilon + \nabla p_\varepsilon = f_\varepsilon, \ \operatorname{div} y_\varepsilon = 0 & (t,x) \in (0,T) \times \Omega, \\ y_\varepsilon(t,0,x_2) = 0 & (t,x_2) \in (0,T) \times (0,1), \\ y_\varepsilon(0,x) = y_0, \quad y_\varepsilon(T,x) = 0 & x \in \Omega. \end{cases} \tag{4.62}$$

The sequence of the functions $y_\varepsilon$ can be constructed in the following way: First let us choose a sufficiently small number $\delta = \delta(\varepsilon) > 0$ such that

$$\|f\|_{L^{p_0}(T-3\delta,T;V')} \leq \varepsilon/10.$$

• On the interval between $t = 0$ and $t = T - 3\delta$, we do not exert any control. So in this interval our function $y_\varepsilon$ is given by the solution to the Navier–Stokes system with homogeneous Dirichlet boundary conditions.

• Next, on the interval $[T-3\delta, T-2\delta]$, we consider a function $\tilde{y}_{0,\varepsilon} \in V \cap C_0^\infty(\Omega)$ close to $y(T-3\delta, x)$ in $V$. In particular,

$$\|\tilde{y}_{0,\varepsilon} - y(T-3\delta, \cdot)\|_V \le \delta^3.$$

On the interval $[T-3\delta, T-2\delta]$ we set

$$y_\varepsilon(t,x) = \frac{(t-T+3\delta)}{\delta}\tilde{y}_{0,\varepsilon}(x) - \frac{(t-T+2\delta)}{\delta}y(T-3\delta, x),$$
$$(t,x) \in [T-3\delta, T-2\delta] \times \Omega.$$

• As the next step, on the segment $[T-2\delta, T-2\delta+2/\tilde{N}]$, we look for the solution $u_\varepsilon$ in the form

$$y_\varepsilon(t,x) = \tilde{N}^2\tilde{U}(t,x) + \boldsymbol{y}(t,x) - \tilde{V}(t,x), \quad p_\varepsilon(t,x) = \tilde{r}(t,x),$$

where $\tilde{U}(t,x) = U(t-T+2\delta, x)$, $\boldsymbol{y}(t,x) = \tilde{y}(t-T+2\delta, x)$, $\tilde{V}(t,x) = \theta(t-T+2\delta)V(t-T+2\delta, x)$, $\tilde{r}(t,x) = \theta(t)r(t-T+2\delta, x)$.

The function $\tilde{y}$ solves the following controllability problem for the transport equation:

$$\begin{cases} \partial_t\tilde{y} + \tilde{N}^2(U,\nabla)\tilde{y} + \tilde{N}^2(\tilde{y},\nabla)U = 0 & (t,x) \in (0,T)\times\tilde{K}, \\ \tilde{y}(t,0,x_2) = 0 & (t,x_2) \in (0,T)\times\mathbb{R}^1, \\ \tilde{y}(0,x) = \tilde{y}_{0,\varepsilon}, \ \ \tilde{y}(1/\tilde{N},x) = 0 & x \in \tilde{K}. \end{cases} \tag{4.63}$$

The function $\tilde{V}$ is a correction term, which ensure that the vector field $y_\varepsilon$ is divergence free:

$$\begin{cases} \partial_t V - \Delta V = \nabla r, \ \ \operatorname{div} V = \operatorname{div}\tilde{y} & (t,x) \in (0,T)\times\tilde{K}, \\ V(t,0,x_2) = V(t,1,x_2) = 0 & (t,x_2) \in (0,T)\times\mathbb{R}^1, \\ V(t,x_1,x_2) = V(t,x_1,x_2+2) & (t,x_1,x_2) \in (0,T)\times(0,2)\times\mathbb{R}^1, \\ V(0,x) = 0 & x \in \tilde{K}. \end{cases} \tag{4.64}$$

There exists a positive constant $C > 0$ independent of $\tilde{N}$ such that

$$\|V\|_{C([0,2/\tilde{N}];L^2(\tilde{K}))} + \|V_{x_2}\|_{C([0,2/\tilde{N}];L^2(\tilde{K}))} \le \frac{C}{\tilde{N}^{\frac{1}{8}}}. \tag{4.65}$$

This estimate is the consequence of the global version of sharp regularity result for the pressure obtained in [10]. Finally $\theta = \theta(t) \in C^2([0, 2/\tilde{N}])$ is an arbitrary function such that

$$\theta(t) = 1, \ \ t \in [0, 1/\tilde{N}], \quad \text{and} \quad \theta(t) = 0 \ \text{ in a neighborhood of } 2/\tilde{N}.$$

Let $\boldsymbol{y}_\varepsilon = 0$ for $(t, x) \in (T - 2\delta, T - 2\delta + 2/\tilde{N}) \times \Omega$. We set $f_\varepsilon = \partial_t y_\varepsilon - \Delta y_\varepsilon + (y_\varepsilon, \nabla) y_\varepsilon$ for all $(t, x) \in [T - 2\delta, T - 2\delta + 2/\tilde{N}] \times \Omega$.

A short computation and (4.65) imply

$$\|\tilde{f}_\varepsilon\|_{L^{p_0}(T-2\delta, T-2\delta+2/\tilde{N}; V')} \leq C\tilde{N}^{7/8-1/p_0}.$$

Thanks to our choice of $p_0$, this constant tends to zero as $\tilde{N} \to +\infty$.

• Finally, on the interval $[T - 2\delta + 1/\tilde{N}, T]$, we take $f_\varepsilon \equiv 0$ and we try to find a boundary control which drives the associated solution of (4.62) which starts at time $t = T - 2\delta + 2/\tilde{N}$ from the initial condition $\tilde{N}^2 U(2/\tilde{N}, x)$ to zero at time $t = T$.

Observe that we have $y_\varepsilon(T - 2\delta + 2/\tilde{N}, x) = \tilde{N}^2 U(2/\tilde{N}, x)$ since $\theta(2/\tilde{N}) = 0$.

By Theorem 2.1 for any $\overline{z}_0 \in L^2(0, 1)$, there exists a boundary control $\rho = \rho(t) \in L^2(0, 2/\tilde{N} - 2\delta)$ such that the solution of

$$\begin{cases} \partial_t \overline{z} - \partial^2_{x_1 x_1} \overline{z} = 0 & (t, x_1) \in (0, T) \times (0, 1), \\ \overline{z}(t, 0) = 0, \ \ \overline{z}(t, 1) = \rho(t) & t \in (0, T), \\ \overline{z}(0, x_1) = \overline{z}_0 & x_1 \in (0, 1). \end{cases}$$

satisfies

$$\overline{z}(2\delta - 2/\tilde{N}, x_1) = 0, \quad x_1 \in (0, 1).$$

Then it suffices to take

$$y_\varepsilon(t, x) = (0, \overline{z}(t - T + 2\delta - 2/\tilde{N}, x_1)), \quad (t, x) \in (T - 2\delta + 2/\tilde{N}, T) \times \Omega,$$

with $\overline{z}$ the solution of the previous null controllability problem with initial condition

$$\overline{z}_0(x_1) = z(2/\tilde{N}, x_1) \quad x_1 \in (0, 1).$$

The construction of the function $y_\varepsilon$ is finished.

## References


[1] Belishev, M. I., On approximating properties of Solutions of the heat equation. In *Control theory of partial differential equations*, Lect. Notes Pure Appl. Math. 242, Chapman & Hall/CRC, Boca Raton, FL, 2005, 43–50.

[2] Bardos, C., Lebeau, G., Rauch, J., Sharp sufficient conditions for the observation, control, and stabilization of wave from boundaries. *SIAM J. Control Optim.* **30** (1992), 1024–1065.

[3] Coron, J.-M., On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier-Slip boundary conditions. *ESAIM Contrôle Optim. Calc. Var.* **1** (1996), 35–75.

[4] Coron, J.-M., On the controllability of 2-D incompressible perfect fluids. *J. Math. Pures Appl.* **75** (1996), 155–188.

[5] Coron, J.-M., Contrôlabilité exacte frontière de l'équation d'Euler des fluids parfaits incompresibles bidimensionnesls. *C. R. Acad. Sci. Paris Sér. I Math.* **317** (1993), 271–276.

[6] Coron, J.-M., On null assymtotic stabilization of the 2-D Euler equation of incompressible fluids on simply connected domains. *SIAM J. Control Optim.* **37** (1999), 1874–1896.

[7] Coron, J.-M., Fursikov, A. V., Global exact controllability of the 2-D Navier-Stokes equations on manifold without boundary. *Russian J. Math. Phys.* **4** (1996), 1–20.

[8] Egorov, Y. V., Some problems in the theory of optimal control. *Ž. Vyčisl. Mat. i Mat. Fiz.* **3** (1963), 887–904.

[9] Fabre, C., Puel, J.-P., Zuazua, E., Approximate controllability of the semilinear heat equation. *Proc. Roy. Soc. Edinburgh Sect. A* **125** (1995), 31–61.

[10] Fabre, C., Lebeau, G., Prolongement unique des solutions de l'équation de Stokes. *Comm. Partial Differential Equations* **21** (1996), 573–596.

[11] Fattorini, H. O., Boundary control of temperature distributions in a parallepipedon. *SIAM J. Control Optim.* **13** (1975), 1–13.

[12] Fernandez-Cara, E., Guerrero, S., Imanuvilov, O., Puel, J.-P., Local exact controllability of Stokes and Navier-Stokes system. *J. Math. Pures Appl.* **83** (2005), N12, 1501–1542.

[13] Fernandez-Cara, E., Guerrero, S., Imanuvilov, O., Puel, J.-P., Some controllability results for the N-dimentional Navier-stokes and Boussinesque system with N-1 scalar controls. *SIAM J. Control Optim.*, to appear.

[14] Fursikov, A. V., Imanuvilov, O. Yu., Local exact controllability of two dimensional Navier-Stokes system with control on the part of the boundary. *Sb. Math.* **187** (1996), 1355–1390.

[15] Fursikov, A. V., Imanuvilov, O. Yu., Local exact boundary controllability of the Boussinesq equation. *SIAM J. Control Optim.* **36** (1988), 391–421.

[16] Fursikov, A. V., Imanuvilov, O. Yu., *Controllability of evolution equations*. Lecture Notes Ser. 34, Seoul National University, Seoul 1996.

[17] Fursikov, A. V., Imanuvilov, O. Yu., Exact controllability of the Navier-Stokes equations and the Boussinesq system. *Russian Math. Surveys* **54** (1999), 565–618.

[18] Glass, O., Contrôlabilité de l'équation d'Euler tridimensionelle pour les fluides parfaits incompressibles. *Séminaire sur les Équations aux Dérivees Partielles*, 1997–1998, Exp. No XV, 11 pp, École Polytechnique, Palaiseau 1998.

[19] Glass, O., Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles en dimension 3. *C. R. Acad. Sci. Paris Sér. I Math.* **325** (1997), 987–992.

[20] Guerrero, S., Imanuvilov, O. Yu., Puel, J. P., Remarks on global approximate controllability for the 2-D Navier-Stokes system with Dirichlet boundary conditions. Submitted.

[21] Imanuvilov, O. Yu., On exact controllability for the Navier-Stokes equations. *ESAIM Contrôle Optim. Calc. Var.* **3** (1998), 97–131.

[22] Imanuvilov, O. Yu., Boundary controllability of parabolic equations. *Sb. Math.* **186** (1995), 879–900.

[23] Imanuvilov, O. Yu., Remarks on exact controllability for Navier-Stokes equations. *ESAIM Contrôle Optim. Calc. Var.* **6** (2001), 39–72.

[24] Imanuvilov, O. Yu., Puel, J. P., Global Carleman estimates for weak solutions of elliptic nonhomogeneous Dirichlet problems. *Internat. Math. Res. Notices* **2003** (16) (2003), 883–913.

[25] Imanuvilov, O. Yu., Yamamoto, M., Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations. *Publ. Res. Inst. Math. Sci.* **39** (2) (2003), 227–274.

[26] Kazemi, M. V., Klibanov, M. A., Stability estimates for ill-posed Cauchy problems involving hyperbolic equations and inequalities. *Appl. Anal.* **50** (1993), 93–102.

[27] Komornik, V., *Exact controllability and stabilization.* RAM Res. Appl. Math., Masson, Paris 1994.

[28] Lagnese, J. E., Lions, J.-L., *Modelling, Analysis and Control of Thin Plates*. Rech. Math. Appl. 6, Masson, Paris 1988.

[29] Lasiecka, I., Triggiani, R., Yao, P. F., Inverse observability estimates for second order hyperbolic equations with variable coefficients. *J. Math. Anal. Appl.* **235** (1999), 13–57.

[30] Lebeau, G., Robbiano, L., Contrôle exact de l'équation de la chaleur. *Comm. Partial Differential Equations* **20** (1995), 336–356.

[31] Lions, J.-L., *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués.* Tome 1, Rech. Math. Appl. 8, Masson, Paris 1988.

[32] Lions, J.-L., Are there connections between turbulence and controllability? 9$^{e}$ Conférence internationale de l'INRIA, Antibes, 12–15 juin 1990.

[33] Lin Guo, Y.-J., Littman, W., Null boundary controllability for semilinear heat equations. *Appl. Math. Optim.* **32** (1995), 281–316.

[34] Littman, W., Boundary control theory for hyperbolic and parabolic equations with constant coefficients. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) (1978), 567–580.

[35] Russell, D. L., Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. *SIAM Rev.* **20** (1978), 639–739.

[36] Russell, D. L., A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. *Studies in Appl. Math.* **52** (1973), 189–212.

[37] Tataru, D., Boundary controllability for conservative PDE's. *Appl. Math. Optim.* **31** (1995), 257–296.

[38] Tataru, D., A priori estimates of Carleman's type in domains with boundary. *J. Math. Pure Appl.* **73** (1994), 355–387.

[39] Triggiani, R., Yao, P. F., Carleman estimates with no lower order terms for general Riemannian waves equations. Global uniqueness and observability in one shot. *Appl. Math. Optim.* **46** (2002), 331–375.

Department of Mathematics, Colorado State University, 101 Werber Building, Fort Collins, CO 80523-1874, U.S.A.

E-mail: oleg@math.colostate.edu

# Port-Hamiltonian systems: an introductory survey

Arjan van der Schaft

**Abstract.** The theory of port-Hamiltonian systems provides a framework for the geometric description of network models of physical systems. It turns out that port-based network models of physical systems immediately lend themselves to a Hamiltonian description. While the usual geometric approach to Hamiltonian systems is based on the canonical symplectic structure of the phase space or on a Poisson structure that is obtained by (symmetry) reduction of the phase space, in the case of a port-Hamiltonian system the geometric structure derives from the *interconnection* of its sub-systems. This motivates to consider Dirac structures instead of Poisson structures, since this notion enables one to define Hamiltonian systems with algebraic constraints. As a result, any power-conserving interconnection of port-Hamiltonian systems again defines a port-Hamiltonian system.

The port-Hamiltonian description offers a systematic framework for analysis, control and simulation of complex physical systems, for lumped-parameter as well as for distributed-parameter models.

**Mathematics Subject Classification (2000).** Primary 93A30,70H05,70H45,70Q05, 70G45, 93B29, 37J60; Secondary 93C10, 93C15, 93C20, 37K05.

**Keywords.** Interconnection, Dirac structures, constrained systems, Hamiltonian DAEs, stabilization, boundary control, conservation laws.

## 1. Introduction

Historically, the Hamiltonian approach has its roots in analytical mechanics and starts from the principle of least action, and proceeds, via the Euler-Lagrange equations and the Legendre transform, towards the Hamiltonian equations of motion. On the other hand, the *network* approach stems from electrical engineering, and constitutes a cornerstone of mathematical systems theory. While most of the *analysis* of physical systems has been performed within the Lagrangian and Hamiltonian framework, the network point of view is prevailing in *modelling* and *simulation* of (complex) physical engineering systems.

The framework of port-Hamiltonian systems *combines* both points of view, by associating with the interconnection structure of the network model a *geometric structure* given by a (pseudo-) *Poisson structure*, or more generally a *Dirac structure*. The Hamiltonian dynamics is then defined with respect to this Dirac structure *and* the Hamiltonian given by the total stored energy. Furthermore, port-Hamiltonian systems are *open* dynamical systems, which interact with their environment through ports. Re-

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

sistive effects are included by terminating some of these ports on energy-dissipating elements.

Dirac structures encompass the geometric structures which are classically being used in the geometrization of mechanics (that is, Poisson structures and pre-symplectic structures), and allow to describe the geometric structure of dynamical systems with algebraic *constraints*. Furthermore, Dirac structures allow to extend the Hamiltonian description of *distributed-parameter systems* to include variable boundary conditions, leading to distributed-parameter port-Hamiltonian systems with boundary ports.

**Acknowledgements.** This survey is based on joint work with several co-authors. In particular I thank Bernhard Maschke and Romeo Ortega for fruitful collaborations.

## 2. Finite-dimensional port-Hamiltonian systems

In this section we recapitulate the basics of finite-dimensional port-Hamiltonian systems. For more details we refer e.g. to [19], [17], [20], [33], [34], [30], [36], [12], [5].

**2.1. From classical Hamiltonian equations to port-Hamiltonian systems.** The standard *Hamiltonian equations* for a mechanical system are given as

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p}(q, p), \\ \dot{p} &= -\frac{\partial H}{\partial q}(q, p) + F \end{aligned} \tag{1}$$

where the *Hamiltonian* $H(q, p)$ is the total energy of the system, $q = (q_1, \ldots, q_k)^T$ are generalized configuration coordinates for the mechanical system with $k$ degrees of freedom, $p = (p_1, \ldots, p_k)^T$ is the vector of generalized momenta, and the input $F$ is the vector of external generalized forces. The state space of (1) with local coordinates $(q, p)$ is called the *phase space*.

One immediately derives the following *energy balance*:

$$\frac{d}{dt}H = \frac{\partial^T H}{\partial q}(q, p)\dot{q} + \frac{\partial^T H}{\partial p}(q, p)\dot{p} = \frac{\partial^T H}{\partial p}(q, p)F = \dot{q}^T F, \tag{2}$$

expressing that the increase in energy of the system is equal to the supplied work (*conservation of energy*). This motivates to define the *output* of the system as $e = \dot{q}$ (the vector of generalized velocities).

System (1) is more generally given in the following form

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p}(q,p), \quad (q,p) = (q_1,\dots,q_k,p_1,\dots,p_k), \\ \dot{p} &= -\frac{\partial H}{\partial q}(q,p) + B(q)f, \qquad f \in \mathbb{R}^m, \\ e &= B^T(q)\frac{\partial H}{\partial p}(q,p) \quad (= B^T(q)\dot{q}), \quad e \in \mathbb{R}^m, \end{aligned} \tag{3}$$

with $B(q)f$ denoting the generalized forces resulting from the input $f \in \mathbb{R}^m$. In case $m < k$ we speak of an *underactuated* system. Similarly to (2) we obtain the energy balance

$$\frac{dH}{dt}(q(t),p(t)) = e^T(t)f(t). \tag{4}$$

A further generalization is to consider systems which are described in local coordinates as

$$\begin{aligned} \dot{x} &= J(x)\frac{\partial H}{\partial x}(x) + g(x)f, \qquad x \in \mathcal{X},\ f \in \mathbb{R}^m, \\ e &= g^T(x)\frac{\partial H}{\partial x}(x), \qquad e \in \mathbb{R}^m, \end{aligned} \tag{5}$$

where $J(x)$ is an $n \times n$ matrix with entries depending smoothly on $x$, which is assumed to be *skew-symmetric*, that is $J(x) = -J^T(x)$, and $x = (x_1,\dots,x_n)$ are local coordinates for an $n$-dimensional state space manifold $\mathcal{X}$ (not necessarily even-dimensional as above). Because of skew-symmetry of $J$ we easily recover the energy-balance $\frac{dH}{dt}(x(t)) = e^T(t)f(t)$. We call (5) a *port-Hamiltonian system* with *structure matrix* $J(x)$, input matrix $g(x)$, and *Hamiltonian* $H$ ([17], [19], [18]).

**Remark 2.1.** In many examples the structure matrix $J$ will additionally satisfy an *integrability* condition (the Jacobi-identity) allowing us to find by Darboux's theorem "canonical coordinates". In this case $J$ is the structure matrix of a *Poisson structure* on $\mathcal{X}$.

**Example 2.2.** An important class of systems that naturally can be written as port-Hamiltonian systems, is constituted by mechanical systems with *kinematic constraints* [22]. Consider a mechanical system locally described by $k$ configuration variables $q = (q_1,\dots,q_k)$. Suppose that there are constraints on the generalized velocities $\dot{q}$, described as

$$A^T(q)\dot{q} = 0, \tag{6}$$

with $A(q)$ an $r \times k$ matrix of rank $r$ everywhere. The constraints (6) are called *holonomic* if it is possible to find new configuration coordinates $\bar{q} = (\bar{q}_1,\dots,\bar{q}_k)$ such that the constraints are equivalently expressed as $\dot{\bar{q}}_{k-r+1} = \dot{\bar{q}}_{n-r+2} = \cdots = \dot{\bar{q}}_k = 0$, in which case the kinematic constraints integrate to the *geometric* constraints

$$\bar{q}_{k-r+1} = c_{k-r+1},\dots,\bar{q}_k = c_k \tag{7}$$

for certain constants $c_{k-r+1}, \ldots, c_k$ determined by the initial conditions. Then the system reduces to an *unconstrained* system in the remaining configuration coordinates $(\bar{q}_1, \ldots, \bar{q}_{k-r})$. If it is *not* possible to integrate the kinematic constraints as above, then the constraints are called *nonholonomic*. The equations of motion for the mechanical system with constraints (6) are given by the *constrained Hamiltonian equations*

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p}(q, p), \\ \dot{p} &= -\frac{\partial H}{\partial q}(q, p) + A(q)\lambda + B(q) f, \\ e &= B^T(q)\frac{\partial H}{\partial p}(q, p), \\ 0 &= A^T(q)\frac{\partial H}{\partial p}(q, p). \end{aligned} \tag{8}$$

The *constrained* state space is therefore given as the following subset of the phase space:

$$\mathcal{X}_c = \left\{(q, p) \mid A^T(q)\frac{\partial H}{\partial p}(q, p) = 0\right\}. \tag{9}$$

One way of proceeding is to *eliminate* the constraint forces, and to *reduce* the equations of motion to the constrained state space, leading (see [32] for details) to a port-Hamiltonian system (5). The structure matrix of this reduced port-Hamiltonian system satisfies the Jacobi identity if and only if the constraints (6) are *holonomic* [32]. An alternative way of approaching the system (8) is to formalize it directly as an *implicit* port-Hamiltonian system (with respect to a Dirac structure), as will be the topic of Section 2.3.

**2.2. From port-based network modelling to port-Hamiltonian systems.** In this subsection we take a different point of view by emphasizing how port-Hamiltonian systems directly arise from *port-based network models* of physical systems.

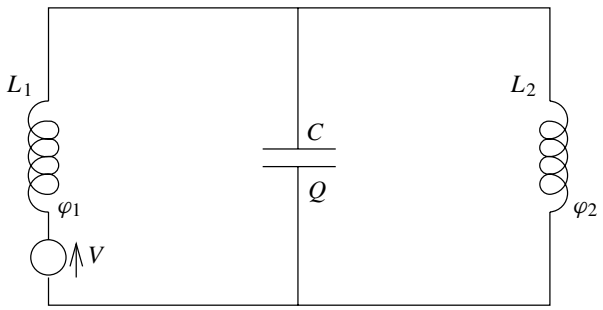



Figure 1. Controlled LC-circuit.

In network models of complex physical systems the overall system is regarded as the *interconnection* of energy-storing elements via basic interconnection (balance) laws such as Newton's third law or Kirchhoff's laws, as well as power-conserving elements like transformers, kinematic pairs and ideal constraints, together with energy-dissipating elements [3], [14], [13]. The basic point of departure for the theory of port-Hamiltonian systems is to formalize the basic interconnection laws together with the power-conserving elements by a *geometric structure*, and to define the Hamiltonian as the total energy stored in the system. This is already illustrated by the following simple example.

**Example 2.3** (LCTG circuits). Consider a controlled LC-circuit (see Figure 1) consisting of two inductors with magnetic energies $H_1(\varphi_1)$, $H_2(\varphi_2)$ ($\varphi_1$ and $\varphi_2$ being the magnetic flux linkages), and a capacitor with electric energy $H_3(Q)$ ($Q$ being the charge). If the elements are linear then $H_1(\varphi_1) = \frac{1}{2L_1}\varphi_1^2$, $H_2(\varphi_2) = \frac{1}{2L_2}\varphi_2^2$ and $H_3(Q) = \frac{1}{2C}Q^2$. Furthermore let $V = u$ denote a voltage source. Using Kirchhoff's laws one obtains the dynamical equations

$$\begin{aligned} \begin{bmatrix} \dot{Q} \\ \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{bmatrix} &= \underbrace{\begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{J} \begin{bmatrix} \frac{\partial H}{\partial Q} \\ \frac{\partial H}{\partial \varphi_1} \\ \frac{\partial H}{\partial \varphi_2} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u, \\ y &= \frac{\partial H}{\partial \varphi_1} \qquad (= \text{current through voltage source}) \end{aligned} \tag{10}$$

with $H(Q, \varphi_1, \varphi_2) := H_1(\varphi_1) + H_2(\varphi_2) + H_3(Q)$ the total energy. Clearly (by Tellegen's theorem) the matrix $J$ is skew-symmetric.

In this way every LC-circuit with independent elements can be modelled as a port-Hamiltonian system. Similarly any LCTG-circuit with independent elements can be modelled as a port-Hamiltonian system, with $J$ now being determined by Kirchhoff's laws *and* the constitutive relations of the transformers $T$ and gyrators $G$.

**2.3. Dirac structures and implicit port-Hamiltonian systems.** From a general modeling point of view physical systems are, at least in first instance, often described as DAE's, that is, a mixed set of differential and *algebraic* equations. This stems from the fact that in network modeling the system under consideration is regarded as obtained from interconnecting simpler sub-systems. These interconnections usually give rise to algebraic constraints between the state space variables of the sub-systems; thus leading to implicit systems. Therefore it is important to extend the framework of port-Hamiltonian systems to the context of *implicit systems*; that is, systems with algebraic constraints.

**2.3.1. Dirac structures.** In order to give the definition of an implicit port-Hamiltonian system we introduce the notion of a Dirac structure, formalizing the concept of

a power-conserving interconnection, and generalizing the notion of a structure matrix $J(x)$ as encountered before.

Let $\mathcal{F}$ be an $\ell$-dimensional linear space, and denote its dual (the space of linear functions on $\mathcal{F}$) by $\mathcal{F}^*$. The product space $\mathcal{F} \times \mathcal{F}^*$ is considered to be the space of power variables, with power defined by

$$P = \langle f^* | f \rangle, \quad (f, f^*) \in \mathcal{F} \times \mathcal{F}^*, \tag{11}$$

where $\langle f^* | f \rangle$ denotes the duality product. Often we call $\mathcal{F}$ the space of *flows* $f$, and $\mathcal{F}^*$ the space of *efforts* $e$, with the power of an element $(f, e) \in \mathcal{F} \times \mathcal{F}^*$ denoted as $\langle e | f \rangle$.

**Example 2.4.** Let $\mathcal{F}$ be the space of generalized *velocities*, and $\mathcal{F}^*$ be the space of generalized *forces*, then $\langle e | f \rangle$ is mechanical power. Similarly, let $\mathcal{F}$ be the space of *currents*, and $\mathcal{F}^*$ be the space of *voltages*, then $\langle e | f \rangle$ is electrical power.

There exists on $\mathcal{F} \times \mathcal{F}^*$ the canonically defined symmetric bilinear form

$$\langle (f_1, e_1), (f_2, e_2) \rangle_{\mathcal{F} \times \mathcal{F}^*} := \langle e_1 | f_2 \rangle + \langle e_2 | f_1 \rangle \tag{12}$$

for $f_i \in \mathcal{F}$, $e_i \in \mathcal{F}^*$, $i = 1, 2$.

**Definition 2.5** ([6], [8], [7]). A constant Dirac structure on $\mathcal{F}$ is a linear subspace $\mathcal{D} \subset \mathcal{F} \times \mathcal{F}^*$ such that

$$\mathcal{D} = \mathcal{D}^{\perp} \tag{13}$$

where $^{\perp}$ denotes the orthogonal complement with respect to the bilinear form $\langle \, , \rangle_{\mathcal{F} \times \mathcal{F}^*}$.

It immediately follows that the dimension of any Dirac structure $\mathcal{D}$ on an $\ell$-dimensional linear space is equal to $\ell$. Furthermore, let $(f, e) \in \mathcal{D} = \mathcal{D}^{\perp}$. Then by (12)

$$0 = \langle (f, e), (f, e) \rangle_{\mathcal{F} \times \mathcal{F}^*} = 2 \langle e \mid f \rangle. \tag{14}$$

Thus for all $(f, e) \in \mathcal{D}$ we obtain $\langle e \mid f \rangle = 0$. Hence a Dirac structure $\mathcal{D}$ on $\mathcal{F}$ defines a power-conserving relation between the power variables $(f, e) \in \mathcal{F} \times \mathcal{F}^*$, which moreover has maximal dimension.

**Remark 2.6.** For many systems, especially those with 3-D mechanical components, the Dirac structure is actually *modulated* by the energy or geometric variables. Furthermore, the state space $\mathcal{X}$ is a *manifold* and the flows $f_S = -\dot{x}$ corresponding to energy-storage are elements of the tangent space $T_x\mathcal{X}$ at the state $x \in \mathcal{X}$, while the efforts $e_S$ are elements of the co-tangent space $T_x^*\mathcal{X}$.

Modulated Dirac structures often arise as a result of *kinematic constraints*. In many cases, these constraints will be configuration dependent, causing the Dirac structure to be modulated by the configuration variables, cf. Section 2.2.

In general, a port-Hamiltonian system can be represented as in Figure 2. The port variables entering the Dirac structure $\mathcal{D}$ have been split in different parts. First, there are two *internal* ports. One, denoted by $\mathcal{S}$, is corresponding to energy-storage and the other one, denoted by $\mathcal{R}$, is corresponding to internal energy-dissipation (resistive elements). Second, two *external* ports are distinguished. The external port denoted by $\mathcal{C}$ is the port that is accessible for controller action. Also the presence of *sources* may be included in this port. Finally, the external port denoted by $\mathcal{I}$ is the interaction port, defining the interaction of the system with (the rest of) its environment.

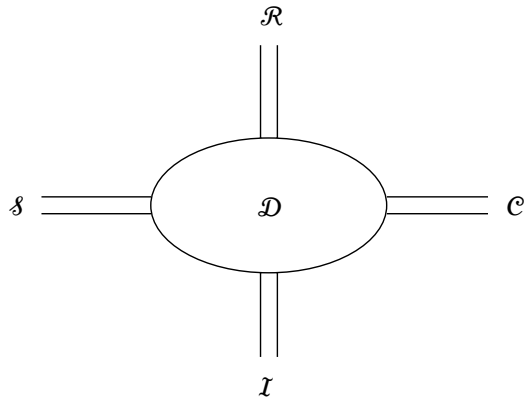



Figure 2. Port-Hamiltonian system.

**2.3.2. Energy storage port.** The port variables associated with the internal storage port will be denoted by $(f_S, e_S)$. They are interconnected to the energy storage of the system which is defined by a finite-dimensional state space manifold $\mathcal{X}$ with coordinates $x$, together with a Hamiltonian function $H\colon \mathcal{X} \to \mathbb{R}$ denoting the energy. The flow variables of the energy storage are given by the *rate* $\dot{x}$ of the energy variables $x$. Furthermore, the effort variables of the energy storage are given by the *co-energy* variables $\frac{\partial H}{\partial x}(x)$, resulting in the energy balance

$$\frac{d}{dt}H = \left\langle \frac{\partial H}{\partial x}(x) \;\middle|\; \dot{x} \right\rangle = \frac{\partial^T H}{\partial x}(x)\dot{x}. \tag{15}$$

(Here we adopt the convention that $\frac{\partial H}{\partial x}(x)$ denotes the *column* vector of partial derivatives of $H$.)

The interconnection of the energy storing elements to the storage port of the Dirac structure is accomplished by setting

$$\begin{aligned} f_S &= -\dot{x}, \\ e_S &= \frac{\partial H}{\partial x}(x). \end{aligned} \tag{16}$$

Hence the energy balance (15) can be also written as

$$\frac{d}{dt}H = \frac{\partial^T H}{\partial x}(x)\dot{x} = -e_S^T f_S. \tag{17}$$

**2.3.3. Resistive port.** The second internal port corresponds to internal energy dissipation (due to friction, resistance, etc.), and its port variables are denoted by $(f_R, e_R)$. These port variables are terminated on a static resistive relation $\mathcal{R}$. In general, a static resistive relation will be of the form

$$R(f_R, e_R) = 0, \tag{18}$$

with the property that for all $(f_R, e_R)$ satisfying (18)

$$\langle e_R \mid f_R \rangle \leq 0. \tag{19}$$

In many cases we may restrict ourselves to *linear* resistive relations. This means that the resistive port variables $(f_R, e_R)$ satisfy linear relations of the form

$$R_f f_R + R_e e_R = 0. \tag{20}$$

The inequality (19) corresponds to the square matrices $R_f$ and $R_e$ satisfying the properties of symmetry and semi-positive definiteness

$$R_f R_e^T = R_e R_f^T \geq 0, \tag{21}$$

together with the dimensionality condition $\text{rank}[R_f | R_e] = \dim f_R$.

Without the presence of additional external ports, the Dirac structure of the port-Hamiltonian system satisfies the power-balance $e_S^T f_S + e_R^T f_R = 0$ which leads to

$$\frac{d}{dt}H = -e_S^T f_S = e_R^T f_R \leq 0. \tag{22}$$

An important special case of resistive relations between $f_R$ and $e_R$ occurs when the resistive relations can be expressed as an *input-output* mapping $f_R = -F(e_R)$, where the resistive characteristic $F: \mathbb{R}^{m_r} \to \mathbb{R}^{m_r}$ satisfies

$$e_R^T F(e_R) \geq 0, \quad e_R \in \mathbb{R}^{m_r}. \tag{23}$$

For *linear* resistive elements this specializes to $f_R = -\tilde{R} e_R$, for some positive semi-definite symmetric matric $\tilde{R} = \tilde{R}^T \geq 0$.

**2.3.4. External ports.** Now, let us consider in more detail the *external* ports to the system. We distinguish between two types of external ports. One is the *control port* $\mathcal{C}$, with port variables $(f_C, e_C)$, which are the port variables which are accessible for controller action. Other type of external port is the *interaction port* $\mathcal{I}$, which denotes the interaction of the port-Hamiltonian system with its environment. The

port variables corresponding to the interaction port are denoted by $(f_I, e_I)$. By taking both the external ports into account the power-balance extends to

$$e_S^T f_S + e_R^T f_R + e_C^T f_C + e_I^T f_I = 0 \tag{24}$$

whereby (22) extends to

$$\frac{d}{dt}H = e_R^T f_R + e_C^T f_C + e_I^T f_I. \tag{25}$$

**2.3.5. Port-Hamiltonian dynamics.** The port-Hamiltonian system with state space $\mathcal{X}$, Hamiltonian $H$ corresponding to the energy storage port $\mathcal{S}$, resistive port $\mathcal{R}$, control port $\mathcal{C}$, interconnection port $\mathcal{I}$, and total Dirac structure $\mathcal{D}$ will be succinctly denoted by $\Sigma = (\mathcal{X}, H, \mathcal{R}, \mathcal{C}, \mathcal{I}, \mathcal{D})$. The dynamics of the port-Hamiltonian system is specified by considering the constraints on the various port variables imposed by the Dirac structure, that is

$$(f_S, e_S, f_R, e_R, f_C, e_C, f_I, e_I) \in \mathcal{D},$$

and to substitute in these relations the equalities $f_S = -\dot{x}$, $e_S = \frac{\partial H}{\partial x}(x)$. This leads to the implicitly defined dynamics

$$\Big( -\dot{x}(t), \frac{\partial H}{\partial x}(x(t)), f_R(t), e_R(t), f_C, (t), e_C(t), f_I(t), e_I(t) \Big) \in \mathcal{D} \tag{26}$$

with $f_R(t), e_R(t)$ satisfying for all $t$ the resistive relation (18):

$$R(f_R, e_R) = 0. \tag{27}$$

In many cases of interest the dynamics (26) will constrain the allowed states $x$, depending on the values of the external port variables $(f_C, e_C)$ and $(f_I, e_I)$. Thus in an equational representation port-Hamiltonian systems generally will consist of a mixed set of *differential* and *algebraic* equations (DAEs).

**Example 2.7** (General LC- circuits). Consider an LC-circuit with general network topology. Kirchhoff's current and voltage laws take the general form

$$A_L{}^T I_L + A_C{}^T I_C + A_P{}^T I_P = 0,$$

$$V_L = A_L \lambda, \quad V_C = A_C \lambda, \quad V_P = A_P \lambda$$

for some matrices $A_L$, $A_C$, $A_S$. Here $I_L$, $I_C$, $I_P$ denote the currents, respectively through the inductors, capacitors and external ports. Likewise, $V_L$, $V_C$, $V_P$ denote the voltages over the inductors, capacitors and external ports. Kirchhoff's current and voltage laws define a Dirac structure between the flows and efforts:

$$f = (I_C, V_L, I_P) = (-\dot{Q}, -\dot{\phi}, I_P),$$
$$e = (V_C, I_L, V_P) = \left( \frac{\partial H}{\partial Q}, \frac{\partial H}{\partial \phi}, V_P \right)$$

with Hamiltonian $H(\phi, Q)$ the total energy. This leads the to port-Hamiltonian system in implicit form

$$\begin{aligned} -\dot{\phi} &= A_L \lambda, \\ \frac{\partial H}{\partial Q} &= A_C \lambda, \\ V_P &= A_P \lambda, \\ 0 &= {A_L}^T \frac{\partial H}{\partial \phi} - {A_C}^T \dot{Q} + {A_P}^T I_P. \end{aligned}$$

**Example 2.8** (Electro-mechanical system). Consider the dynamics of an iron ball in the magnetic field of a controlled inductor: The port-Hamiltonian description of this

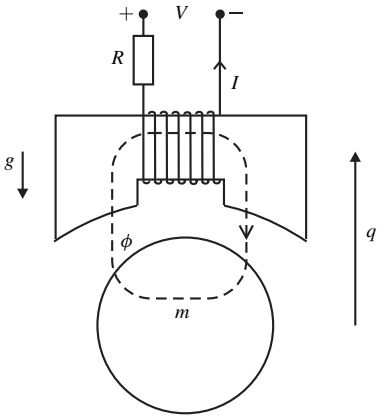



Figure 3. Magnetically levitated ball.

system (with $q$ the height of the ball, $p$ the vertical momentum, and $\varphi$ the magnetic flux of the inductor) is given as

$$\begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \\ \dot{\varphi} \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -\frac{1}{R} \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial q} \\ \frac{\partial H}{\partial p} \\ \frac{\partial H}{\partial \varphi} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} V, \\ I &= \frac{\partial H}{\partial \varphi}. \end{aligned} \tag{28}$$

This is a typical example of a system where the *coupling* between two different physical domains (mechanical and magnetic) takes place via the Hamiltonian

$$H(q, p, \varphi) = mgq + \frac{p^2}{2m} + \frac{\varphi^2}{2k_1(1 - \frac{q}{k_2})}$$

where the last term depends both on a magnetic variable (in this case $\varphi$) and a mechanical variable (in this case the height $q$).

**2.4. Input-state-output port-Hamiltonian systems.** An important special case of port-Hamiltonian systems is the class of *input-state-output port-Hamiltonian systems*, where there are no algebraic constraints on the state space variables, and the flow and effort variables of the resistive, control and interaction port are split into conjugated input–output pairs. Input–state–output port-Hamiltonian systems without interaction port are of the form

$$\begin{aligned} \dot{x} &= [J(x) - R(x)]\frac{\partial H}{\partial x}(x) + g(x)u, \\ y &= g^T(x)\frac{\partial H}{\partial x}(x) \end{aligned} \tag{29}$$

where $u$, $y$ are the input–output pairs corresponding to the control port $\mathcal{C}$. Here the matrix $J(x)$ is skew-symmetric, while the matrix $R(x) = R^T(x) \geq 0$ specifies the resistive structure, and is given as $R(x) = g_R^T(x)\tilde{R}g_R(x)$ for some linear resistive relation $f_R = -\tilde{R}e_R$, $\tilde{R} = \tilde{R}^T \geq 0$, with $g_R$ representing the input matrix corresponding to the resistive port. The underlying Dirac structure of the system is then given by the graph of the skew-symmetric linear map

$$\begin{pmatrix} -J(x) & -g_R(x) & -g(x) \\ g_R^T(x) & 0 & 0 \\ g^T(x) & 0 & 0 \end{pmatrix}. \tag{30}$$

## 3. Control by interconnection of port-Hamiltonian systems

The basic property of port-Hamiltonian systems is that the power-conserving interconnection of any number of port-Hamiltonian systems is again a port-Hamiltonian system.

To be explicit, consider two port-Hamiltonian systems $\Sigma_A$ and $\Sigma_B$ with Dirac structures $\mathcal{D}_A$ and $\mathcal{D}_B$ and Hamiltonians $H_A$ and $H_B$, defined on state spaces $\mathcal{X}_A$, respectively $\mathcal{X}_B$. For convenience, split the ports of the Dirac structures $\mathcal{D}_A$ and $\mathcal{D}_B$ into the internal energy storage ports and all remaining external ports whose port-variables are denoted respectively by $f_A$, $e_A$ and $f_B$, $e_B$. Now, consider any *interconnection* Dirac structure $\mathcal{D}_I$ involving the port-variables $f_A$, $e_A$, $f_B$, $e_B$ possibly together with additional port-variables $f_I$, $e_I$. Then the interconnection of the systems $\Sigma_A$ and $\Sigma_B$ via $\mathcal{D}_I$ is again a port-Hamiltonian system with respect to the composed Dirac structure $\mathcal{D}_A \circ \mathcal{D}_I \circ \mathcal{D}_B$, involving as port-variables the internal storage port-variables of $\mathcal{D}_A$ and $\mathcal{D}_B$ together with the additional port-variables $F_I$, $e_I$. For details we refer to [5], [34], [30].

Furthermore, the state space of the interconnected port-Hamiltonian system is the product of the two state spaces $\mathcal{X}_A \times \mathcal{X}_B$, while its Hamiltonian is simply the sum $H_A + H_B$ of the two Hamiltonians.

This basic statement naturally extends to the interconnection of any number of port-Hamiltonian systems via an interconnection Dirac structure.

*Control by port-interconnection* is based on designing a controller system which is *interconnected* to the control port with port-variables $(f_C, e_C)$. *In principle* this implies that we only consider *collocated* control, where the controller will only use the information about the plant port-Hamiltonian system that is contained in the conjugated pairs $(f_C, e_C)$ of port variables of the control port, without using additional information about the plant (e.g. corresponding to observation on other parts of the plant system). In the second place, we will restrict attention to controller systems which are themselves *also* port-Hamiltonian systems. There are two main reasons for this. One is that by doing so the closed-loop system is *again* a port-Hamiltonian system, allowing to easily ensure some desired properties. Furthermore, it will turn out that the port-Hamiltonian framework suggests useful ways to construct port-Hamiltonian controller systems. Second reason is that port-Hamiltonian controller systems allow in principle for a physical system realization (thus linking to passive control and systems design) and physical *interpretation* of the controller action.

Since we do not know the environment (or only have very limited information about it), but on the other hand, the system *will* interact with this unknown environment, the task of the controller is often two-fold: 1) to achieve a desired control goal (e.g. set-point regulation or tracking) if the interaction with the environment is marginal or can be compensated, 2) to make sure that the controlled system has a desired interaction behavior with its environment. It is fair to say that up to now the development of the theory of control of port-Hamiltonian systems has mostly concentrated on the second aspect (which at the same time, is often underdeveloped in other control theories).

Most successful approaches to deal with the second aspect of the control goal are those based on the concept of "*passivity*", such as *dissipativity theory* [38], *impedance control* [13] and *Intrinsically Passive Control* (IPC) [36]. In fact, the port-Hamiltonian control theory can be regarded as an enhancement to the theory of passivity, making a much closer link with complex physical systems modeling at one hand and with the theory of dynamical systems (in particular, Hamiltonian dynamics) at the other hand.

As said above, we will throughout consider controller systems which are again port-Hamiltonian systems. We will use the same symbols as above for the internal and external ports and port-variables of the controller port-Hamiltonian system, with an added overbar $\bar{}$ or a superscript $^c$ in order to distinguish it from the plant system. (The interaction port of the controller system may be thought of as an extra possibility for additional controller action (outer-loop control).) In order to further distinguish the plant system and the controller we denote the state space of the plant system by $\mathcal{X}_p$ with coordinates $x_p$, the Dirac structure by $\mathcal{D}_p$ and its Hamiltonian by $H_p$, while we will denote the state space manifold of the controller system by $\mathcal{X}_c$ with coordinates $x_c$, its Dirac structure by $\mathcal{D}_c$ and its Hamiltonian by $H_c \colon \mathcal{X}_c \to \mathbb{R}$. The interconnection

of the plant port-Hamiltonian system with the controller port-Hamiltonian system is obtained by equalizing the port variables at the control port by

$$\begin{aligned} f_C &= -\bar{f}_C, \\ e_C &= \bar{e}_C \end{aligned} \tag{31}$$

where $\bar{f}_C$, $\bar{e}_C$ denote the control port variables of the controller system. Here, the minus sign is inserted to have a uniform notion of direction of power flow. Clearly, this 'synchronizing' interconnection is power-conserving, that is $e_C^T f_C + \bar{e}_C^T \bar{f}_C = 0$.

**Remark 3.1.** A sometimes useful alternative is the *gyrating* power-conserving interconnection

$$\begin{aligned} f_C &= -\bar{e}_C, \\ e_C &= \bar{f}_C. \end{aligned} \tag{32}$$

In fact, the standard feedback interconnection can be regarded to be of this type.

For both interconnection constraints it directly follows from the theory of composition of Dirac structures that the interconnected (closed-loop) system is again a port-Hamiltonian system with Dirac structure determined by the Dirac structures of the plant PH system and the controller PH system.

The resulting interconnected PH system has state space $\mathcal{X}_p \times \mathcal{X}_c$, Hamiltonian $H_p + H_c$, resistive ports $(f_R, e_R, \bar{f}_R, \bar{e}_R)$ and interaction ports $(f_I, e_I, \bar{f}_I, \bar{e}_I)$, satisfying the power-balance

$$\frac{d}{dt}(H_p + H_c) = e_R^T f_R + \bar{e}_R^T \bar{f}_R + e_I^T f_I + \bar{e}_I^T \bar{f}_I \leq e_I^T f_I + \bar{e}_I^T \bar{f}_I \tag{33}$$

since both $e_R^T f_R \leq 0$ and $\bar{e}_R^T \bar{f}_R \leq 0$. Hence we immediately recover the state space formulation of the passivity theorem, see e.g. [31], if $H_p$ and $H_c$ are both non-negative, implying that the plant and the controller system are passive (with respect to their controller and interaction ports and storage functions $H_p$ and $H_c$), then also the closed -loop system is passive (with respect to the interaction ports and storage function $H_p + H_c$.)

Furthermore, due to the Hamiltonian structure, we can go *beyond* the passivity theorem, and we can derive conditions which ensure that we can passify and/or stabilize plant port-Hamiltonian systems for which the Hamiltonian $H_p$ does *not* have a minimum at the desired equilibrium.

**3.1. Stabilization by Casimir generation.** What does the power-balance (33)mean for the stability properties of the closed-loop system, and how can we design the controller port-Hamiltonian system in such a way that the closed-loop system has desired stability properties? Let us first consider the stability of an arbitrary port-Hamiltonian system $\Sigma = (\mathcal{X}, H, \mathcal{R}, \mathcal{C}, \mathcal{I}, \mathcal{D})$ *without* control or interaction ports,

that is, an autonomous port-Hamiltonian system $\Sigma = (\mathcal{X}, H, \mathcal{R}, \mathcal{D})$. Clearly, the power-balance (33) reduces to

$$\frac{d}{dt}H = e_R^T f_R \leq 0. \tag{34}$$

Hence we immediately infer by standard Lyapunov theory that if $x^*$ is a minimum of the Hamiltonian $H$ then it will be a *stable* equilibrium of the autonomous port-Hamiltonian system $\Sigma = (\mathcal{X}, H, \mathcal{R}, \mathcal{D})$, which is actually *asymptotically stable* if the dissipation term $e_R^T f_R$ is negative *definite* outside $x^*$, or alternatively if some sort of detectability condition is satisfied, guaranteeing asymptotic stability by the use of LaSalle's Invariance principle (see for details e.g. [31]).

However, what can we say if $x^*$ is *not* a minimum of $H$, and thus we cannot directly use $H$ as a Lyapunov function?

A well-known method in Hamiltonian systems, sometimes called the Energy-Casimir method, is to use in the Lyapunov analysis next to the Hamiltonian *other* conserved quantities (dynamical invariants) which may be present in the system. Indeed, if we may find other conserved quantities then candidate Lyapunov functions can be sought within the class of *combinations* of the Hamiltonian $H$ and those conserved quantities. In particular, if we can find a conserved quantity $C\colon \mathcal{X} \to \mathbb{R}$ such that $V := H + C$ has a minimum at the desired equilibrium $x^*$ then we can still infer stability or asymptotic stability by replacing (34) by

$$\frac{d}{dt}V = e_R^T f_R \leq 0 \tag{35}$$

and thus using $V$ as a Lyapunov function.

For the application of the Energy-Casimir method one may distinguish between two main cases. First situation occurs if the desired equilibrium $x^*$ is not a stationary point of $H$, and one looks for a conserved quantity $C$ such that $H + C$ has a minimum at $x^*$. This for example happens in the case that the desired set-point $x^*$ is *not* an equilibrium of the uncontrolled system, but only a controlled equilibrium of the system. Second situation occurs when $x^*$ *is* a stationary point of $H$, but not a minimum.

Functions that are conserved quantities of the system for *every* Hamiltonian are called *Casimir functions* or simply Casimirs. Casimirs are completely characterized by the Dirac structure of the port-Hamiltonian system. Indeed, a function $C\colon \mathcal{X} \to \mathbb{R}$ is a Casimir function of the autonomous port-Hamiltonian system (without energy dissipation) $\Sigma = (\mathcal{X}, H, \mathcal{D})$ if and only if the gradient vector $e = \frac{\partial^T C}{\partial x}$ satisfies

$$e^T f_S = 0 \quad \text{for all } f_S \text{ for which there exists } e_S \text{ such that } (f_S, e_S) \in \mathcal{D}. \tag{36}$$

Indeed, (36) is equivalent to

$$\frac{d}{dt}C = \frac{\partial^T C}{\partial x}(x(t))\dot{x}(t) = \frac{\partial^T C}{\partial x}(x(t))\, f_S = e^T f_S = 0 \tag{37}$$

for every port-Hamiltonian system $(\mathcal{X}, H, \mathcal{D})$ with the same Dirac structure $\mathcal{D}$. By the generalized skew-symmetry of the Dirac structure (36) is equivalent to the requirement that $e = \frac{\partial^T C}{\partial x}$ satisfies

$$(0, e) \in \mathcal{D}.$$

Similarly, we define a Casimir function for a port-Hamiltonian system with dissipation $\Sigma = (\mathcal{X}, H, \mathcal{R}, \mathcal{D})$ to be any function $C\colon \mathcal{X} \to \mathbb{R}$ satisfying

$$(0, e, 0, 0) \in \mathcal{D}. \tag{38}$$

Indeed, this will imply that

$$\frac{d}{dt}C = \frac{\partial^T C}{\partial x}(x(t))\dot{x}(t) = \frac{\partial^T C}{\partial x}(x(t))\, f_p = e^T f_p = 0 \tag{39}$$

for every port-Hamiltonian system $(\mathcal{X}, H, \mathcal{R}, \mathcal{D})$ with the same Dirac structure $\mathcal{D}$. (In fact by definiteness of the resistive structures the satisfaction of (39) for a particular resistive structure $\mathcal{R}$ *implies* the satisfaction for *all* resistive structures $\mathcal{R}$.)

Now let us come back to the design of a controller port-Hamiltonian system such that the closed-loop system has desired stability properties. Suppose we want to stabilize the plant port-Hamiltonian system $(\mathcal{X}_p, H_p, \mathcal{R}, \mathcal{C}, \mathcal{D}_p)$ around a desired equilibrium $x_p^*$. We know that for every controller port-Hamiltonian system the closed-loop system satisfies

$$\frac{d}{dt}(H_p + H_c) = e_R^T f_R + \bar{e}_R^T \bar{f}_R \leq 0. \tag{40}$$

What if $x^*$ is not a minimum for $H_p$? A possible strategy is to *generate* Casimir functions $C(x_p, x_c)$ for the closed-loop system by choosing the controller port-Hamiltonian system in an appropriate way. Thereby we generate candidate Lyapunov functions for the closed-loop system of the form

$$V(x_p, x_c) := H_p(x_p) + H_c(x_c) + C(x_p, x_c)$$

where the controller Hamiltonian function $H_c\colon \mathcal{X}_c \to \mathbb{R}$ still has to be designed. The goal is thus to construct a function $V$ as above in such a way that $V$ has a minimum at $(x_p^*, x_c^*)$ where $x_c^*$ still remains to be chosen. This strategy thus is based on finding all the achievable closed-loop Casimirs. Furthermore, since the closed-loop Casimirs are based on the closed-loop Dirac structures, this reduces to finding all the achievable closed-loop Dirac structures $\mathcal{D} \circ \bar{\mathcal{D}}$.

*Another* way to interpret the generation of Casimirs for the closed-loop system is to look at the level sets of the Casimirs as *invariant submanifolds* of the combined plant and controller state space $\mathcal{X}_p \times \mathcal{X}_c$. Restricted to every such invariant submanifold (part of) the controller state can be expressed as a function of the plant state, whence the closed-loop Hamiltonian restricted to such an invariant manifold can be seen as a

*shaped* version of the plant Hamiltonian. To be explicit (see e.g. [31], [24], [25] for details) suppose that we have found Casimirs of the form

$$x_{ci} - F_i(x_p), \quad i = 1, \ldots, n_p$$

where $n_p$ is the dimension of the controller state space, then on every invariant manifold $x_{ci} - F_i(x_p) = \alpha_i$, $i = 1, \ldots, n_p$, where $\alpha = (\alpha_1, \ldots, \alpha_{n_p})$ is a vector of constants depending on the initial plant and controller state, the closed-loop Hamiltonian can be written as

$$H_s(x_p) := H_p(x_p) + H_c(F(x_p) + \alpha),$$

where, as before, the controller Hamiltonian $H_c$ still can be assigned. This can be regarded as *shaping* the original plant Hamiltonian $H_p$ to a new Hamiltonian $H_s$.

**3.2. Port Control.** In broad terms, the *Port Control* problem is to design, given the plant port-Hamiltonian system, a controller port-Hamiltonian system such that the *behavior* at the interaction port of the plant port-Hamiltonian system is a desired one, or close to a desired one. This means that by adding the controller system we seek to shape the external behavior at the interaction port of the plant system. If the desired external behavior at this interaction port is given in input–output form as a desired (dynamic) impedance, then this amounts to the Impedance Control problem as introduced and studied by Hogan and co-workers [13]; see also [36] for subsequent developments.

The Port Control problem, as stated in this generality, immediately leads to two fundamental questions: 1). Given the plant PH system, and the controller PH system to be arbitrarily designed, what are the achievable behaviors of the closed-loop system at the interaction port of the plant? 2). If the desired behavior at the interaction port of the plant is not achievable, then what is the closest achievable behavior? Of course, the second question leaves much room for interpretation, since there is no obvious interpretation of what we mean by 'closest behavior'. Also the first question in its full generality is not easy to answer, and we shall only address an important subproblem.

An obvious observation is that the desired behavior, in order to be achievable, needs to be the port behavior of a PH system. This leads already to the problem of characterizing those external behaviors which are port behaviors of port-Hamiltonian systems. Secondly, the Port Control problem can be split into a number of subproblems. Indeed, we know that the closed-loop system arising from interconnection of the plant PH system with the controller PH system is specified by a Hamiltonian which is just the sum of the plant Hamiltonian and the controller Hamiltonian, and a resistive structure which is the "product" of the resistive structure of the plant and of the controller system, together with a Dirac structure which is the *composition* of the plant Dirac structure and the controller Dirac structure. Therefore an important subproblem is again to characterize the *achievable closed-loop Dirac structures.* On the other hand, a fundamental problem in addressing the Port Control problem in general theoretical terms is the lack of a systematic way to specify 'desired behavior'.

The problem of Port Control is to determine the controller system in such a way that the port behavior in the port variables $f_I, e_I$ is a desired one. In this particular (simple and linear) example the desired behavior can be quantified e.g. in terms of a desired stiffness and damping of the closed-loop system, which is easily expressed in terms of the closed-loop transfer function from $f_I$ to $e_I$. Of course, on top of the requirements on the closed-loop transfer function we would also require internal stability of the closed-loop system. For an appealing example of port control of port-Hamiltonian systems within a context of hydraulic systems we refer to [15].

**3.3. Energy Control.** Consider two port-Hamiltonian systems $\Sigma_i$ (without internal dissipation) in input–state–output form

$$\begin{aligned} \dot{x}_i &= J_i(x_i)\frac{\partial H_i}{\partial x_i} + g_i(x_i)u_i, \\ y_i &= g_i^T(x_i)\frac{\partial H_i}{\partial x_i}, \qquad i = 1, 2, \end{aligned} \tag{41}$$

both satisfying the power-balance $\frac{d}{dt}H_i = y_i^T u_i$. Suppose now that we want to transfer the energy from the port-Hamiltonian system $\Sigma_1$ to the port-Hamiltonian system $\Sigma_2$, while keeping the total energy $H_1 + H_2$ constant. This can be done by using the following output feedback

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 & -y_1 y_2^T \\ y_2 y_1^T & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{42}$$

Since the matrix in (42) is skew-symmetric it immediately follows that the closed-loop system composed of systems $\Sigma_1$ and $\Sigma_2$ linked by the power-conserving feedback is energy-preserving, that is $\frac{d}{dt}(H_1 + H_2) = 0$. However, if we consider the individual energies then we notice that

$$\frac{d}{dt}H_1 = -y_1^T y_1 y_2^T y_2 = -\|y_1\|^2\|y_2\|^2 \leq 0 \tag{43}$$

implying that $H_1$ is decreasing as long as $\|y_1\|$ and $\|y_2\|$ are different from 0. Conversely, as expected since the total energy is constant,

$$\frac{d}{dt}H_2 = y_2^T y_2 y_1^T y_1 = \|y_2\|^2\|y_1\|^2 \geq 0 \tag{44}$$

implying that $H_2$ is increasing at the same rate. In particular, if $H_1$ has a minimum at the zero equilibrium, and $\Sigma_1$ is zero-state observable, then all the energy $H_1$ of $\Sigma_1$ will be transferred to $\Sigma_2$, provided that $\|y_2\|$ is not identically zero (which again can be guaranteed by assuming that $H_2$ has a minimum at the zero equilibrium, and that $\Sigma_2$ is zero-state observable).

If there is internal energy dissipation, then this energy transfer mechanism still works. However, the fact that $H_2$ grows or not will depend on the balance between the energy delivered by $\Sigma_1$ to $\Sigma_2$ and the internal loss of energy in $\Sigma_2$ due to dissipation.

We conclude that this particular scheme of power-conserving energy transfer is accomplished by a skew-symmetric output feedback, which is *modulated* by the values of the output vectors of both systems. Of course this raises, among others, the question of the efficiency of the proposed energy-transfer scheme, and the need for a systematic quest of similar power-conserving energy-transfer schemes. We refer to [9] for a similar but different energy-transfer scheme directly motivated by the structure of the example (control of a snakeboard).

**3.4. Achievable closed-loop Dirac structures.** In all the control problems discussed above the basic question comes up what are the achievable closed-loop Dirac structures based on a given plant Dirac structure and a controller Dirac structure, which still is to be determined.

**Theorem 3.2** ([5]). *Given any plant Dirac structure $\mathcal{D}_p$, a certain interconnected $\mathcal{D} = \mathcal{D}_p \circ \mathcal{D}_c$ can be achieved by a proper choice of the controller Dirac structure $\mathcal{D}_c$ if and only if the following two equivalent conditions are satisfied:*

$$\mathcal{D}_p^0 \subset \mathcal{D}^0,$$
$$\mathcal{D}^\pi \subset \mathcal{D}_p^\pi$$

*where*

$$\mathcal{D}_p^0 := \{f_1, e_1) \mid (f_1, e_1, 0, 0) \in \mathcal{D}_p\},$$
$$\mathcal{D}_p^\pi := \{(f_1, e_1) \mid \textit{there exists } (f_2^P, e_2^P) \textit{ with } (f_1, e_1, f_2^P, e_2^P) \in \mathcal{D}_p\},$$
$$\mathcal{D}^0 := \{(f_1, e_1) \mid (f_1, e_1, 0, 0) \in \mathcal{D}\},$$
$$\mathcal{D}^\pi := \{(f_1, e_1) \mid \textit{there exists } (f_3, e_3) \textit{ with } (f_1, e_1, f_3, e_3) \in \mathcal{D}\}.$$

An important application of the above theorem concerns the characterization of Casimir functions which can be achieved by interconnecting a given plant port-Hamiltonian system with a controller port-Hamiltonian system.

## 4. Distributed-parameter port-Hamiltonian systems

The treatment of infinite-dimensional Hamiltonian systems in the literature is mostly confined to systems with boundary conditions such that the energy exchange through the boundary is *zero*. On the other hand, in many applications the interaction with the environment (e.g. actuation or measurement) will actually take place through the boundary of the system. In [35] a framework has been developed to represent classes of physical distributed-parameter systems with boundary energy flow as *infinite-dimensional port-Hamiltonian systems*. It turns out that in order to allow the inclusion of boundary variables in distributed-parameter systems the concept of (an

infinite-dimensional) Dirac structure provides again the right type of generalization with respect to the existing framework [23] using Poisson structures.

As we will discuss in the next three examples, the port-Hamiltonian formulation of distributed-parameter systems is closely related to the general framework for describing basic distributed-parameter systems as systems of conservation laws, see e.g. [11], [37].

**Example 4.1** (Inviscid Burger's equation). The viscous *Burger's equation* is a scalar parabolic equation defined on a one-dimensional spatial domain (interval) $Z = [a, b] \subset \mathbb{R}$, with the state variable $\alpha(t, z) \in \mathbb{R}$, $z \in Z$, $t \in I$, where $I$ is an interval of $\mathbb{R}$, satisfying the partial differential equation

$$\frac{\partial \alpha}{\partial t} + \alpha \frac{\partial \alpha}{\partial z} - \nu \frac{\partial^2 \alpha}{\partial z^2} = 0. \tag{45}$$

The *inviscid* ($\nu = 0$) Burger's equations may be alternatively expressed as

$$\frac{\partial \alpha}{\partial t} + \frac{\partial}{\partial z} \beta = 0 \tag{46}$$

where the state variable $\alpha(t, z)$ is called the *conserved quantity* and the function $\beta := \frac{\alpha^2}{2}$ the *flux* variable. Eq. (46) is called a *conservation law*, since by integration one obtains the *balance equation*

$$\frac{d}{dt} \int_a^b \alpha \, dz = \beta(a) - \beta(b). \tag{47}$$

Furthermore, according to the framework of Irreversible Thermodynamics [27], one may express the flux $\beta$ as a function of the *generating force* which is the *variational derivative* of some functional $H(\alpha)$ of the state variable. The variational derivative $\frac{\delta H}{\delta \alpha}$ of a functional $H(\alpha)$ is uniquely defined by the requirement

$$H(\alpha + \varepsilon \eta) = H(\alpha) + \varepsilon \int_a^b \frac{\delta H}{\delta \alpha} \eta \, dz + O(\varepsilon^2) \tag{48}$$

for any $\varepsilon \in \mathbb{R}$ and any smooth function $\eta(z, t)$ such that $\alpha + \varepsilon \eta$ satisfies the same boundary conditions as $\alpha$ [23]. For the inviscid Burger's equation one has $\beta = \frac{\delta H}{\delta \alpha}$, where

$$H(\alpha) = \int_a^b \frac{\alpha^3}{6} \, dz. \tag{49}$$

Hence the inviscid Burger's equation may be also expressed as

$$\frac{\partial \alpha}{\partial t} = -\frac{\partial}{\partial z} \frac{\delta H}{\delta \alpha}. \tag{50}$$

This defines an infinite-dimensional Hamiltonian system in the sense of [23] with respect to the skew-symmetric operator $\frac{\partial}{\partial z}$ that is defined on the functions with support contained in the interior of the interval $Z$.

From this formulation one derives that the Hamiltonian $H(\alpha)$ is *another* conserved quantity. Indeed, by integration by parts

$$\frac{d}{dt}H = \int_a^b \frac{\delta H}{\delta \alpha} \cdot -\frac{\partial}{\partial z}\frac{\delta H}{\delta \alpha} dz = \frac{1}{2}\left(\beta^2(a) - \beta^2(b)\right). \tag{51}$$

We note that the right-hand side is a *function of the flux variables* evaluated at the boundary of the spatial domain $Z$.

The second example consists of a system of *two* conservations laws, corresponding to the case of two physical domains in interaction.

**Example 4.2** (The p-system, cf. [11], [37]). The p-system is a model for e.g. a one-dimensional gas dynamics. Again, the spatial variable $z$ belongs to an interval $Z \subset \mathbb{R}$, while the dependent variables are the specific volume $v(t, z) \in \mathbb{R}^+$, the velocity $u(t, z)$ and the pressure functional $p(v)$ (which for instance in the case of an ideal gas with constant entropy is given by $p(v) = Av^{-\gamma}$ where $\gamma \geq 1$). The *p-system* is then defined by the following system of partial differential equations

$$\begin{aligned} \frac{\partial v}{\partial t} - \frac{\partial u}{\partial z} &= 0, \\ \frac{\partial u}{\partial t} + \frac{\partial\, p(v)}{\partial z} &= 0 \end{aligned} \tag{52}$$

representing respectively conservation of mass and of momentum. By defining the state vector as $\alpha(t, z) = (v, u)^T$, and the vector-valued flux $\beta(t, z) = (-u, p(v))^T$ the p-system is rewritten as

$$\frac{\partial \alpha}{\partial t} + \frac{\partial}{\partial z}\beta = 0. \tag{53}$$

Again, according to the framework of Irreversible Thermodynamics, the flux vector may be written as function of the variational derivatives of some functional. Indeed, consider the energy functional $H(\alpha) = \int_a^b \mathcal{H}(v, u)dz$ where the energy density $\mathcal{H}(v, u)$ is given as the sum of the internal energy and the kinetic energy densities

$$\mathcal{H}(v, u) = \mathcal{U}(v) + \frac{u^2}{2} \tag{54}$$

with $-\mathcal{U}(v)$ a primitive function of the pressure. (Note that for simplicity the mass density has been set equal to 1, and hence no difference is made between the velocity and the momentum.) The flux vector $\beta$ may be expressed in terms of the variational derivatives of $H$ as

$$\beta = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\delta H}{\delta v} \\ \frac{\delta H}{\delta u} \end{pmatrix}. \tag{55}$$

The anti-diagonal matrix represents the canonical coupling between two physical domains: the kinetic and the potential (internal) domain. Thus the variational derivative

of the total energy with respect to the state variable of one domain generates the flux variable for the other domain. Combining eqns. (53) and (55), the p-system may thus be written as the Hamiltonian system

$$\begin{pmatrix} \frac{\partial \alpha_1}{\partial t} \\ \frac{\partial \alpha_2}{\partial t} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{\partial}{\partial z} \\ -\frac{\partial}{\partial z} & 0 \end{pmatrix} \begin{pmatrix} \frac{\delta H}{\delta \alpha_1} \\ \frac{\delta H}{\delta \alpha_2} \end{pmatrix}. \tag{56}$$

Using again integration by parts, one may derive the following *energy balance equation*:

$$\frac{d}{dt} H = \beta_1(a)\beta_2(a) - \beta_1(b)\beta_2(b). \tag{57}$$

Notice again that the right-hand side of this power-balance equation is a quadratic function of the fluxes at the boundary of the spatial domain.

The last example is the *vibrating string*. It is again a system of two conservation laws representing the canonical interdomain coupling between the kinetic energy and the elastic potential energy. However in this example the *classical* choice of the state variables leads to express the total energy as a function of some of the *spatial derivatives* of the state variables.

**Example 4.3** (Vibrating string). Consider an elastic string subject to traction forces at its ends, with spatial variable $z \in Z = [a, b] \subset \mathbb{R}$. Denote by $u(t, z)$ the displacement of the string and the velocity by $v(t, z) = \frac{\partial u}{\partial t}$. Using the vector of state variables $x(t, z) = (u, v)^T$, the dynamics of the vibrating string is described by the system of partial differential equations

$$\frac{\partial x}{\partial t} = \begin{pmatrix} v \\ \frac{1}{\mu}\frac{\partial}{\partial z}\left(T\frac{\partial u}{\partial z}\right) \end{pmatrix} \tag{58}$$

where the first equation is simply the definition of the velocity and the second one is Newton's second law. Here $T$ denotes the elasticity modulus, and $\mu$ the mass density. The total energy is $H(x) = U(u) + K(v)$, where the elastic potential energy $U$ is a function of the *strain* $\frac{\partial u}{\partial z}(t, z)$

$$U(u) = \int_a^b \frac{1}{2}\, T \left(\frac{\partial u}{\partial z}\right)^2 dz \tag{59}$$

and the kinetic energy $K$ depends on the velocity $v(t, z) = \frac{\partial u}{\partial t}$ as

$$K(v) = \int_a^b \frac{1}{2}\, \mu v(t, z)^2\, dz. \tag{60}$$

Thus the total system (58) may be expressed as

$$\frac{\partial x}{\partial t} = \begin{pmatrix} 0 & \frac{1}{\mu} \\ -\frac{1}{\mu} & 0 \end{pmatrix} \begin{pmatrix} \frac{\delta H}{\delta u} \\ \frac{\delta H}{\delta v} \end{pmatrix} \tag{61}$$

where $\frac{\delta H}{\delta u} = \frac{\delta U}{\delta u} = -\frac{\partial}{\partial z}\left(T\,\frac{\partial u}{\partial z}\right)$ is the elastic force and $\frac{\delta H}{\delta v} = \frac{\delta K}{\delta v} = \mu v$ is the momentum.

In this formulation, the system is *not* anymore expressed as a system of conservation laws since the time-derivative of the state variables is a function of the variational derivatives of the energy *directly*, and *not* the spatial derivative of a function of the variational derivatives as before. Instead of being a simplification, this reveals a drawback for the case of non-zero energy flow through the boundary of the spatial domain. Indeed, in this case the *variational derivative has to be completed by a boundary term* since the Hamiltonian functional depends on the *spatial derivatives of the state*. For example, in the computation of the variational derivative of the elastic potential energy $U$ one obtains by integration by parts that $U(u+\varepsilon\eta) - U(u)$ equals

$$-\varepsilon \int_a^b \frac{\partial}{\partial z}\left(T\frac{\partial u}{\partial z}\right)\eta\, dz + \varepsilon\left[\eta\left(T\frac{\partial u}{\partial z}\right)\right]_a^b + O(\varepsilon^2) \tag{62}$$

and the second term in this expression constitutes an extra boundary term.

Alternatively we now formulate the vibrating string as a system of two conservation laws. Take as alternative vector of state variables $\alpha(t,z) = (\varepsilon,\, p)^T$, where $\varepsilon$ denotes the *strain* $\alpha_1 = \varepsilon = \frac{\partial u}{\partial z}$ and $p$ denotes the *momentum* $\alpha_2 = p = \mu v$. Recall that in these variables the total energy is written as

$$H_0 = \int_a^b \frac{1}{2}\left(T\alpha_1^2 + \frac{1}{\mu}\alpha_2^2\right) dz \tag{63}$$

and directly depends on the state variables and *not* on their spatial derivatives. Furthermore, one defines the flux variables to be the *stress* $\beta_1 = \frac{\delta H_0}{\delta \alpha_1} = T\alpha_1$ and the *velocity* $\beta_2 = \frac{\delta H_0}{\delta \alpha_1} = \frac{\alpha_2}{\mu}$. In matrix notation, the flux vector $\beta$ is thus expressed as a function of the variational derivatives $\frac{\delta H_0}{\delta \alpha}$ by

$$\beta = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \frac{\delta H_0}{\delta \alpha}. \tag{64}$$

Hence the vibrating string may be alternatively expressed by the system of two conservation laws

$$\frac{\partial \alpha}{\partial t} = \begin{pmatrix} 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & 0 \end{pmatrix} \frac{\delta H_0}{\delta \alpha} \tag{65}$$

satisfying the power balance equation (57).

**4.1. Systems of two conservation laws in interaction.** Let us now consider the *general class* of distributed-parameter systems consisting of two conservation laws with the canonical coupling as in the above examples of the p-system and the vibrating string. Let the spatial domain $Z \subset \mathbb{R}^n$ be an $n$-dimensional smooth manifold with smooth $(n-1)$-dimensional boundary $\partial Z$. Denote by $\Omega^k(Z)$ the vector space of (differential) $k$-forms on $Z$ (respectively by $\Omega^k(\partial Z)$ the vector space of $k$-forms on $\partial Z$).

Denote furthermore by $\Omega = \bigoplus_{k\geq 0} \Omega^k(Z)$ the algebra of differential forms over $Z$ and recall that it is endowed with an exterior product $\wedge$ and an exterior derivation d.

**Definition 4.4.** A *system of conservation laws* is defined by a set of *conserved quantities* $\alpha_i \in \Omega^{k_i}(Z)$, $i \in \{1, \ldots, N\}$ where $N \in \mathbb{N}, k_i \in \mathbb{N}$, defining the state space $\mathcal{X} = \bigotimes_{i=1,\ldots,N} \Omega^{k_i}(Z)$, and satisfying a set of *conservation laws*

$$\frac{\partial \alpha_i}{\partial t} + \mathrm{d}\beta_i = g_i \tag{66}$$

where $\beta_i \in \Omega^{k_i-1}(Z)$ denote the set of *fluxes* and $g_i \in \Omega^{k_i}(Z)$ denote the set of *distributed interaction forms*. In general, the fluxes $\beta_i$ are defined by so-called *closure equations*

$$\beta_i = J(\alpha_i, z), \quad i = 1, \ldots, N \tag{67}$$

leading to a closed form for the dynamics of the conserved quantities $\alpha_i$. The integral form of the conservation laws yields the following *balance equations*

$$\frac{d}{dt}\int_S \alpha_i + \int_{\partial S} \beta_i = \int_S g_i \tag{68}$$

for any surface $S \subset Z$ of dimension equal to the degree of $\alpha_i$.

**Remark 4.5.** A common case is that $Z = \mathbb{R}^3$ and that the conserved quantities are 3-forms, that is, the balance equation is evaluated on volumes of the 3-dimensional space. In this case () takes in vector calculus notation the familiar form

$$\frac{\partial \alpha_i}{\partial t}(z,t) + \mathrm{div}_z \beta_i = g_i, \quad i = 1, \ldots, n. \tag{69}$$

However, systems of conservation laws may correspond to differential forms of any degree. Maxwell's equations are an example where the conserved quantities are differential forms of degree 2.

In the sequel, as in our examples sofar, we consider a particular class of systems of conservation laws where the closure equations are such that fluxes are linear functions of the variational derivatives of the Hamiltonian functional. First recall the general definition of the *variational derivative* of a functional $H(\alpha)$ with respect to the differential form $\alpha \in \Omega^p(Z)$ (generalizing the definition given before).

**Definition 4.6.** Consider a density function $\mathcal{H}\colon \Omega^p(Z) \times Z \to \Omega^n(Z)$ where $p \in \{1, \ldots, n\}$, and denote by $H := \int_Z \mathcal{H} \in \mathbb{R}$ the associated functional. Then the uniquely defined differential form $\frac{\delta H}{\delta \alpha} \in \Omega^{n-p}(Z)$ which satisfies for all $\Delta\alpha \in \Omega^p(Z)$ and $\varepsilon \in \mathbb{R}$

$$H(\alpha + \varepsilon\Delta\alpha) = \int_Z \mathcal{H}(\alpha) + \varepsilon \int_Z \left[\frac{\delta H}{\delta \alpha} \wedge \Delta\alpha\right] + \mathrm{O}\left(\varepsilon^2\right)$$

is called the *variational derivative* of $H$ with respect to $\alpha \in \Omega^p(Z)$.

**Definition 4.7.** *Systems of two conservation laws with canonical interdomain coupling* are systems involving a pair of conserved quantities $\alpha_p \in \Omega^p(Z)$ and $\alpha_q \in \Omega^q(Z)$, differential forms on the $n$-dimensional spatial domain $Z$ of degree $p$ and $q$ respectively, satisfying $p+q = n+1$ ('complementarity of degrees'). The closure equations generated by a *Hamiltonian density function* $\mathcal{H}: \Omega^p(Z) \times \Omega^q(Z) \times Z \to \Omega^n(Z)$ resulting in the Hamiltonian $H := \int_Z \mathcal{H} \in \mathbb{R}$ are given by

$$\begin{pmatrix} \beta_p \\ \beta_q \end{pmatrix} = \varepsilon \begin{pmatrix} 0 & (-1)^r \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\delta H}{\delta \alpha_p} \\ \frac{\delta H}{\delta \alpha_q} \end{pmatrix} \tag{70}$$

where $r = pq+1$, $\varepsilon \in \{-1, +1\}$, depending on the sign convention of the considered physical domain.

Define the vector of *flow variables* to be the time-variation of the state, and the *effort variables* to be the variational derivatives

$$\begin{pmatrix} f_p \\ f_q \end{pmatrix} = \begin{pmatrix} \frac{\partial \alpha_p}{\partial t} \\ \frac{\partial \alpha_q}{\partial t} \end{pmatrix}, \quad \begin{pmatrix} e_p \\ e_q \end{pmatrix} = \begin{pmatrix} \frac{\delta H}{\delta \alpha_p} \\ \frac{\delta H}{\delta \alpha_q} \end{pmatrix}. \tag{71}$$

Their product equals again the time-variation of the Hamiltonian

$$\frac{dH}{dt} = \int_Z (e_p \wedge f_p + e_q \wedge f_q). \tag{72}$$

Using the conservation laws (4.5) for $g_i = 0$, the closure relations (70) and the properties of the exterior derivative $d$ and Stokes' theorem, one obtains

$$\begin{aligned} \frac{dH}{dt} &= \int_Z \varepsilon \beta_q \wedge (-d\beta_p) + (-1)^r \beta_p \wedge \varepsilon(-d\beta_q) \\ &= -\varepsilon \int_Z \beta_q \wedge d\beta_p + (-1)^q \beta_q \wedge d\beta_p = -\varepsilon \int_{\partial Z} \beta_q \wedge \beta_p. \end{aligned} \tag{73}$$

Finally, as before we define the power-conjugated pair of *flow and effort variables on the boundary* as the *restriction* of the flux variables to the boundary $\partial Z$ of the domain $Z$:

$$\begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = \begin{pmatrix} \beta_q|_{\partial Z} \\ \beta_p|_{\partial Z} \end{pmatrix}. \tag{74}$$

On the total space of power-conjugated variables, that is, the differential forms $(f_p, e_p)$ and $(f_q, e_q)$ on the domain $Z$ and the differential forms $(f_\partial, e_\partial)$ defined on the boundary $\partial Z$, one defines an *interconnection structure* by Eqn. (74) together with

$$\begin{pmatrix} f_q \\ f_p \end{pmatrix} = \varepsilon \begin{pmatrix} 0 & (-1)^r d \\ d & 0 \end{pmatrix} \begin{pmatrix} e_q \\ e_p \end{pmatrix}. \tag{75}$$

This interconnection can be formalized as a special type of Dirac structure, called Stokes–Dirac structure, leading to the definition of distributed-parameter port-Hamiltonian systems [35].

## 5. Concluding remarks

We have surveyed some of the recently developed theory of port-Hamiltonian systems; for further applications towards modeling, analysis, simulation and control we refer to the literature cited below.

From the geometric point of view many questions regarding port-Hamiltonian systems are waiting to be investigated. A theory of symmetry and reduction of port-Hamiltonian systems has been explored in [29], [1], while some questions concerning integrability of Dirac structures have been studied in [7]. A main question for distributed-parameter port-Hamiltonian systems concerns the relation with variational calculus.

## References


[1] Blankenstein, G., van der Schaft, A. J., Symmetry and reduction in implicit generalized Hamiltonian systems. *Rep. Math. Phys.* **47** (2001), 57–100.

[2] Bloch, A. M. and Crouch, P. E., Representations of Dirac structures on vector spaces and nonlinear $LC$ circuits. In *Differential geometry and control* (ed. by G. Ferreyra, R. Gardner, H. Hermes, H. Sussmann), Proc. Sympos. Pure Math. 64, Amer. Math. Soc., Providence, RI, 1999, 103–117.

[3] Breedveld, P. C., Physical systems theory in terms of bond graphs. PhD thesis, University of Twente, Faculty of Electrical Engineering, 1984.

[4] R. W. Brockett, Control theory and analytical mechanics. In *Geometric Control Theory* (ed. by C. Martin, R. Hermann), Lie Groups: History, Frontiers and Applications VII, Math. Sci. Press, Brookline 1977, 1–46.

[5] Cervera, J., van der Schaft, A. J., Banos, A., Interconnection of port-Hamiltonian systems and composition of Dirac structures. *Automatica*, submitted.

[6] Courant, T. J., Dirac manifolds. *Trans. Amer. Math. Soc.* **319** (1990), 631–661.

[7] Dalsmo, M., and van der Schaft, A. J., On representations and integrability of mathematical structures in energy-conserving physical systems. *SIAM J. Control Optim.* **37** (1999), 54–91.

[8] Dorfman, I., *Dirac Structures and Integrability of Nonlinear Evolution Equations*. John Wiley, Chichester 1993.

[9] Duindam, V., Blankenstein, G., Stramigioli, S., Port-Based Modeling and Analysis of Snakeboard Locomotion. In *Proceedings 16th International Symposium on Mathematical Theory of Networks and Systems* (MTNS2004), Leuven 2004.

[10] Escobar, G., van der Schaft, A. J., and Ortega, R., A Hamiltonian viewpoint in the modelling of switching power converters. *Automatica* **35** (Special Issue on Hybrid Systems) (1999), 445–452.

[11] Godlewsky, E., and Raviart, P., *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Appl. Math. Sci. 118, Springer-Verlag, New York 1996.

[12] Golo, G., van der Schaft, A. J., Breedveld, P. C., Maschke, B. M., Hamiltonian formulation of bond graphs. In *Nonlinear and Hybrid Systems in Automotive Control* (ed. by R. Johansson, A. Rantzer), Springer-Verlag, London 2003, 351–372.

[13] Hogan, N., Impedance Control: An approach to manipulation. *J. Dyn. Systems, Measurements Control* **107** (1985), 1–24.

[14] Karnopp, D. C., Margolis, D. L., and Rosenberg, R. C., *System Dynamics, A Unified Approach*. John Wiley and Sons, 1990.

[15] Kugi, A., Kemmetmüller, W., Impedance control of Hydraulic piston actuators. In *Proceedings 6th IFAC Symposium on Nonlinear Control Systems* (NOLCOS 2004), 1–3 September, Stuttgart.

[16] Marsden, J. E., and Ratiu, T. S., *Introduction to Mechanics and Symmetry*. Texts Appl. Math. 17, Springer-Verlag, New York 1994.

[17] Maschke, B. M., and van der Schaft, A. J., Port-controlled Hamiltonian systems: Modelling origins and system-theoretic properties. In *Proc. 2nd IFAC NOLCOS, Bordeaux* 1992, 282–288.

[18] Maschke, B. M., and van der Schaft, A. J., System-theoretic properties of port-controlled Hamiltonian systems. In *Systems and Networks: Mathematical Theory and Applications*, Vol. II, Akademie-Verlag, Berlin 1994, 349–352.

[19] Maschke, B. M., van der Schaft, A. J., and Breedveld, P. C., An intrinsic Hamiltonian formulation of network dynamics: non-standard Poisson structures and gyrators. *J. Franklin Inst.* **329** (1992), 923–966.

[20] Maschke, B. M., van der Schaft, A. J., and Breedveld, P. C., An intrinsic Hamiltonian formulation of the dynamics of LC-circuits. *IEEE Trans. Circuits Systems I Fund. Theory Appl.* **42** (1995), 73–82.

[21] Maschke, B. M., Ortega, R., and van der Schaft, A. J., Energy-based Lyapunov functions for forced Hamiltonian systems with dissipation. *IEEE Trans. Automat. Control* **45** (2000), 1498–1502.

[22] Neimark, J. I., and Fufaev, N. A., *Dynamics of Nonholonomic Systems*. Transl. Math. Monogr. 33, Amer. Math. Soc., Providence, RI, 1972.

[23] Olver, P. J., *Applications of Lie Groups to Differential Equations*. Second edition, Grad. Texts in Math. 107, Springer-Verlag, New York 1993.

[24] Ortega, R., van der Schaft, A. J., Maschke, B. M., and Escobar, G., Interconnection and damping assignment passivity-based control of port-controlled Hamiltonian systems. *Automatica* **38** (2002), 585–596.

[25] Ortega, R., van der Schaft, A. J., Mareels, I., and Maschke, B. M., Putting energy back in control. *IEEE Control Syst. Mag.* **21** (2001), 18–33.

[26] Paynter, H. M., *Analysis and design of engineering systems*. M.I.T. Press, Cambridge, MA, 1960.

[27] Prigogine, I., *Introduction to Thermodynamics of Irreversible Processes*. John Wiley and Sons, 1962.

[28] van der Schaft, A. J., *System theoretic properties of physical systems*. CWI Tract 3, Centre for Mathematics and Informatics, Amsterdam 1984.

[29] van der Schaft, A. J., Implicit Hamiltonian systems with symmetry. *Rep. Math. Phys.* **41** (1998), 203–221.

[30] van der Schaft, A. J., Interconnection and geometry. In *The Mathematics of Systems and Control, From Intelligent Control to Behavioral Systems* (ed. by J. W. Polderman, H. L. Trentelman), Groningen 1999, 203–218.

[31] van der Schaft, A. J., *$L_2$-Gain and Passivity Techniques in Nonlinear Control*. 2nd revised and enlarged edition, Comm. Control Engrg. Ser., Springer-Verlag, London 2000 (first edition; Lecture Notes in Control and Inform. Sci. 218, Springer-Verlag, London 1996).

[32] van der Schaft, A. J., and Maschke, B. M., On the Hamiltonian formulation of nonholonomic mechanical systems. *Rep. Math. Phys.* **34** (1994), 225–233.

[33] van der Schaft, A. J., and Maschke, B. M., The Hamiltonian formulation of energy conserving physical systems with external ports. *AEÜ—Arch. Elektron. Übertragungstech.* **49** (1995), 362–371.

[34] van der Schaft, A. J., and Maschke, B. M., Interconnected Mechanica5a Systems, Part I: Geometry of Interconnection and implicit Hamiltonian Systems. In *Modelling and Control of Mechanical Systems* (ed. by A. Astolfi, D. J. N. Limebeer, C. Melchiorri, A. Tornambe, R. B. Vinter), Imperial College Press, London 1997, 1–15.

[35] van der Schaft, A. J., Maschke, B. M., Hamiltonian representation of distributed parameter systems with boundary energy flow. *J. Geom. Phys.* **42** (2002), 166–194.

[36] Stramigioli, S., *Modeling and IPC control of Interactive Mechanical Systems: a coordinate free approach*. Lecture Notes in Control and Inform. Sci. 266, Springer-Verlag, London 2001.

[37] Serre, D., *Systems of Conservation Laws*. Cambridge University Press, Cambridge 1999.

[38] Willems, J. C., Dissipative dynamical systems - Part I: General Theory. *Arch. Rational Mech. Anal.* **45** (1972), 321–351.

Department of Mathematics and Computing Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

E-mail: A.J.van.der.Schaft@math.rug.nl

# Passive linear discrete time-invariant systems

Olof J. Staffans*

**Abstract.** We begin by discussing linear discrete time-invariant i/s/o (input/state/output) systems that satisfy certain 'energy' inequalities. These inequalities involve a quadratic storage function in the state space induced by a positive self-adjoint operator $H$ that may be unbounded and have an unbounded inverse, and also a quadratic supply rate in the combined i/o (input/output) space. The three most commonly studied classes of supply rates are called scattering, impedance, and transmission. Although these three classes resemble each other, we show that there are still significant differences. We then present a new class of s/s (state/signal) systems which have a Hilbert state space and a Kreĭn signal space. The state space is used to store relevant information about the past evolution of the system, and the signal space is used to describe interactions with the surrounding world. A s/s system resembles an i/s/o system apart from the fact that inputs and outputs are not separated from each other. By decomposing the signal space into a direct sum of an input space and an output space one gets a standard i/s/o system, provided the decomposition is *admissible*, and different i/o decompositions lead to different i/o supply rates (for example of scattering, impedance, or transmission type). In the case of non-admissible decompositions we obtain right and left affine representations, both of the s/s system itself, and of the corresponding transfer function. In particular, in the case of a passive system we obtain right and left coprime representations of the generalized transfer functions corresponding to nonadmissible decompositions of the signal space, and we end up with transfer functions which are, e.g., generalized Potapov or Nevanlinna class functions.

**Mathematics Subject Classification (2000).** 93A05, 47A48, 47A67, 47B50.

**Keywords.** Passive, storage function, supply rate, scattering, impedance, transmission, input/state/output, state/signal, Schur function, Carathéodory function, Nevanlinna function, Potapov function, behavior.

## 1. $H$-passive discrete time i/s/o systems

The evolution of a *linear discrete time-invariant i/s/o* (*input/state/output*) *system* $\Sigma_{\mathrm{i/s/o}}$ with a Hilbert *input space* $\mathcal{U}$, a Hilbert *state space* $\mathcal{X}$, and a Hilbert *output space* $\mathcal{Y}$ is described by the system of equations

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n), \\ y(n) &= Cx(n) + Du(n), \quad n \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\}, \\ x(0) &= x_0, \end{aligned} \tag{1.1}$$

*This article is based on recent joint work with Prof. Damir Arov [AS05], [AS06a], [AS06b], [AS06c]. Thank you, Dima, for everything that I have learned from you!

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

where the initial state $x_0 \in \mathcal{X}$ may be chosen arbitrarily and $A\colon \mathcal{X} \to \mathcal{X}$, $B\colon \mathcal{U} \to \mathcal{X}$, $C\colon \mathcal{X} \to \mathcal{Y}$, and $D\colon \mathcal{U} \to \mathcal{Y}$ are bounded linear operators. Equivalently,

$$\begin{bmatrix} x(n+1) \\ y(n) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x(n) \\ u(n) \end{bmatrix}, \quad n \in \mathbb{Z}^+,\ x(0) = x_0, \tag{1.2}$$

where $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{B}\left(\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{U} \end{smallmatrix}\right]; \left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{Y} \end{smallmatrix}\right]\right)$.[1] We call $u = \{u(n)\}_{n=0}^{\infty}$ the *input sequence*, $x = \{x(n)\}_{n=0}^{\infty}$ the *state trajectory*, and $y = \{y(n)\}_{n=0}^{\infty}$ the *output sequence*, and we refer to the triple $(u, x, y)$ as a *trajectory of* $\Sigma_{\mathrm{i/s/o}}$. The operators appearing in (1.1) and (1.2) are usually called as follows: $A$ is the *main operator*, $B$ is the *control operator*, $C$ is the *observation operator*, and $D$ is the *feedthrough operator*. The *transfer function* or *characteristic function* $\mathfrak{D}$ of this system is given by[2]

$$\mathfrak{D}(z) = zC(1_{\mathcal{X}} - zA)^{-1}B + D, \quad z \in \Lambda(A),$$

where $\Lambda(A)$ is the set of points $z \in \mathbb{C}$ for which $1_{\mathcal{X}} - zA$ has a bounded inverse, plus the point at infinity if $A$ has a bounded inverse. Note that $\mathfrak{D}$ is analytic on $\Lambda(A)$, and that $D = \mathfrak{D}(0)$. We shall denote the above system by $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}\right)$. Since all the systems in this paper will be linear and time-invariant and have a discrete time variable we shall in the sequel omit the words "linear discrete time-invariant" and refer to a system of the above type by simply calling it an *i/s/o system*.

The i/s/o system $\Sigma_{\mathrm{i/s/o}}$ is *controllable* if the sets of all states $x(n)$, $n \geq 1$, which appear in some trajectory $(u, x, y)$ of $\Sigma_{\mathrm{i/s/o}}$ with $x_0 = 0$ (i.e., an *externally generated trajectory*) is dense in $\mathcal{X}$. The system $\Sigma_{\mathrm{i/s/o}}$ is *observable* if there do not exist any nontrivial trajectories $(u, x, y)$ where both $u$ and $y$ are identically zero. Finally, $\Sigma_{\mathrm{i/s/o}}$ is *minimal* if $\Sigma_{\mathrm{i/s/o}}$ is both controllable and observable.

In this work we shall primarily be concerned with i/s/o systems which are passive or even conservative. To define these notions we first introduce the notions of a storage function $E_H$ which represents the (internal) energy of the state, and a supply rate $j$ which describes the interchange of energy between the system and its surroundings. In the classical case the *storage* (or *Lyapunov*) *function* $E_H$ is bounded, and it is given by $E_H(x) = \langle x, Hx\rangle_{\mathcal{X}}$, where $H$ is a bounded positive self-adjoint operator on $\mathcal{X}$ (positivity of $H$ means that $\langle x, Hx\rangle_{\mathcal{X}} > 0$ for all $x \neq 0$). However, we shall also consider unbounded storage functions induced by some (possibly unbounded) positive self-adjoint operator $H$ on $\mathcal{X}$. In this case we let $\sqrt{H}$ be the positive self-adjoint square root of $H$, and define the storage function $E_H$ by

$$E_H(x) = \|\sqrt{H}x\|_{\mathcal{X}}^2, \quad x \in \mathcal{D}(\sqrt{H}). \tag{1.3}$$

Clearly, this is equivalent to the earlier definition of $E_H$ if $H$ is bounded. The *supply rate* $j$ will always be a bounded (indefinite) self-adjoint quadratic form on $\mathcal{Y} \oplus \mathcal{U}$,

[1] Here $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{U} \end{smallmatrix}\right]$ is the cartesian product of $\mathcal{X}$ and $\mathcal{U}$, and $\mathcal{B}(\mathcal{U}; \mathcal{Y})$ is the set of bounded linear operators from $\mathcal{U}$ to $\mathcal{Y}$.

[2] $1_{\mathcal{X}}$ is the identity operator in $\mathcal{X}$.

i.e., it can be written in the form

$$j(u, y) = \left\langle \begin{bmatrix} y \\ u \end{bmatrix}, J \begin{bmatrix} y \\ u \end{bmatrix} \right\rangle_{\mathcal{Y} \oplus \mathcal{U}}, \tag{1.4}$$

where $J = \left[\begin{smallmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{smallmatrix}\right]$ is a bounded self-adjoint operator in $\mathcal{Y} \oplus \mathcal{U}$. For simplicity we throughout require $J$ to have a bounded inverse. Often $J$ is taken to be a signature operator (both self-adjoint and unitary), so that $J = J^* = J^{-1}$. In the sequel we shall always use *one and the same supply rate* $j$ for a given system $\Sigma_{\mathrm{i/s/o}}$ and include this supply rate in the notation of the system, thus denoting the system by $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j\right)$ whenever the supply rate is important.

**Definition 1.1.** The i/s/o system $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j\right)$ is *forward $H$-passive*, where $H$ is a positive self-adjoint operator in $\mathcal{X}$, if $x(n) \in \mathcal{D}(\sqrt{H})$ and

$$\|\sqrt{H}x(n+1)\|_{\mathcal{X}}^2 - \|\sqrt{H}x(n)\|_{\mathcal{X}}^2 \le j(u(n), y(n)), \quad n \in \mathbb{Z}^+, \tag{1.5}$$

for every trajectory $(u, x, y)$ of $\Sigma_{\mathrm{i/s/o}}$ with $x_0 \in \mathcal{D}(\sqrt{H})$. If the above inequality holds as an equality then $\Sigma_{\mathrm{i/s/o}}$ is *forward $H$-conservative*.

It is not difficult to see that $\Sigma_{\mathrm{i/s/o}}$ is forward $H$-passive if and only if[3] $H > 0$ is a solution of the (forward) generalized i/s/o *KYP* (*Kalman–Yakubovich–Popov*) inequality[4]

$$\|\sqrt{H}(Ax + Bu)\|_{\mathcal{X}}^2 - \|\sqrt{H}x\|_{\mathcal{X}}^2 \le j(u, Cx + Du), \quad x \in \mathcal{D}(\sqrt{H}),\ u \in \mathcal{U}, \tag{1.6}$$

and that $\Sigma_{\mathrm{i/s/o}}$ is forward $H$-conservative if and only if $H > 0$ is a solution of the corresponding equality. This inequality is named after *Kalman* [Kal63], *Yakubovich* [Yak62], and *Popov* [Pop61] (who at that time restricted themselves to the finite-dimensional case). There is a rich literature on the finite-dimensional version of the KYP inequality and the corresponding equality; see, e.g., [PAJ91], [IW93] and [LR95], and the references mentioned there. In the seventies the classical results on the KYP inequalities were extended to infinite-dimensional systems by V. A. Yakubovich and his students and collaborators (see [Yak74], [Yak75], and [LY76] and the references listed there). There is now also a rich literature on this infinite-dimensional case; see, e.g., the discussion in [Pan99] and the references cited there. However, until recently it was assumed throughout that *either $H$ itself is bounded or $H^{-1}$ is bounded*. The first study of this inequality which permits both $H$ and $H^{-1}$ to be unbounded was done by Arov, Kaashoek and Pik in [AKP05].

Above we have defined *forward $H$-passivity* and forward $H$-conservativity. The corresponding *backward* notions are defined by means of the adjoint i/s/o system

[3] The notation $H > 0$ means that $H$ is a (possibly unbounded) self-adjoint operator satisfying $\langle x, Hx\rangle_{\mathcal{X}} > 0$ for all nonzero $x \in \mathcal{D}(H)$.

[4] In particular, in order for the first term in this inequality to be well-defined we require $A$ to map $\mathcal{D}(\sqrt{H})$ into itself and $B$ to map $\mathcal{U}$ into $\mathcal{D}(\sqrt{H})$.

$\Sigma^*_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A^* & C^* \\ B^* & D^* \end{smallmatrix}\right]; \mathcal{Y}, \mathcal{X}, \mathcal{U}; j_*\right)$ whose trajectories $(y_*, x_*, u_*)$ satisfy the system of equations

$$\begin{aligned} x_*(n+1) &= A^* x_*(n) + C^* y_*(n), \\ u_*(n) &= B^* x_*(n) + D^* y_*(n), \quad n \in \mathbb{Z}^+, \\ x_*(0) &= x_{*0}. \end{aligned} \tag{1.7}$$

Note that this system has the same state space $\mathcal{X}$, but the input and output have been interchanged, so that $\mathcal{Y}$ is the input space and $\mathcal{U}$ is the output space. The appropriate storage function and supply rates for the adjoint system $\Sigma^*_{\mathrm{i/s/o}}$ differ from those of the primal system $\Sigma_{\mathrm{i/s/o}}$: $H$ is replaced by $H^{-1}$, and the primal supply rate $j$ is replaced by the dual supply rate

$$j_*(y_*, u_*) = \left\langle \begin{bmatrix} u_* \\ y_* \end{bmatrix}, J_* \begin{bmatrix} u_* \\ y_* \end{bmatrix} \right\rangle_{\mathcal{U}\oplus\mathcal{Y}}, \tag{1.8}$$

where

$$J_* = \begin{bmatrix} 0 & -1_{\mathcal{U}} \\ 1_{\mathcal{Y}} & 0 \end{bmatrix} J^{-1} \begin{bmatrix} 0 & -1_{\mathcal{Y}} \\ 1_{\mathcal{U}} & 0 \end{bmatrix}. \tag{1.9}$$

**Definition 1.2.** Let $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j\right)$ be an i/s/o system, and let $H$ be a positive self-adjoint operator in $\mathcal{X}$.

(i) $\Sigma_{\mathrm{i/s/o}}$ is *backward $H$-passive* if the adjoint system $\Sigma^*_{\mathrm{i/s/o}}$ is forward $H^{-1}$-passive.

(ii) $\Sigma_{\mathrm{i/s/o}}$ is *backward $H$-conservative* if the adjoint system $\Sigma^*_{\mathrm{i/s/o}}$ is forward $H^{-1}$-conservative.

(iii) $\Sigma_{\mathrm{i/s/o}}$ is $H$-passive if it is both forward and backward $H$-passive.

(iv) $\Sigma_{\mathrm{i/s/o}}$ is $H$-conservative if it is both forward and backward $H$-conservative.

(v) By *passive* or *conservative* (with or without the attributes "forward" or "backward") we mean $1_{\mathcal{X}}$-passive or $1_{\mathcal{X}}$-conservative, respectively.

The generalized KYP inequality for the adjoint i/s/o system $\Sigma^*_{\mathrm{i/s/o}}$ with storage function $E_{H^{-1}}$ is given by[5]

$$\begin{aligned} \|H^{-1/2}(A^* x_* + C^* y_*)\|^2_{\mathcal{X}} - \|H^{-1/2} x_*\|^2_{\mathcal{X}} \le j_*(y_*, B^* x_* + D^* y_*), \\ x_* \in \left(\sqrt{H}\right),\ y_* \in \mathcal{Y}. \end{aligned} \tag{1.10}$$

Thus, $\Sigma_{\mathrm{i/s/o}}$ is backward $H$-passive if and only if $H$ is a solution of (1.10), and $\Sigma_{\mathrm{i/s/o}}$ is backward $H$-conservative if and only if $H$ is a solution of the corresponding equality.

[5] In particular, in order for the first term in this inequality to be well-defined we require $A^*$ to map $\mathcal{R}(\sqrt{H})$ into itself and $C^*$ to map $\mathcal{Y}$ into $\mathcal{R}(\sqrt{H})$.

## 2. Scattering, impedance and transmission supply rates

The three most common supply rates are the following:

(i) The *scattering* supply rate $j_{\mathrm{sca}}(u, y) = -\langle y, y\rangle_{\mathcal{Y}} + \langle u, u\rangle_{\mathcal{U}}$ with signature operator $J_{\mathrm{sca}} = \left[\begin{smallmatrix} -1_{\mathcal{Y}} & 0 \\ 0 & 1_{\mathcal{U}} \end{smallmatrix}\right]$. The signature operator of the dual supply rate is $J_{\mathrm{sca}*} = \left[\begin{smallmatrix} -1_{\mathcal{U}} & 0 \\ 0 & 1_{\mathcal{Y}} \end{smallmatrix}\right]$.

(ii) The *impedance* supply rate $j_{\mathrm{imp}}(u, y) = 2\Re\langle y, \Psi u\rangle_{\mathcal{U}}$ with signature operator $J_{\mathrm{imp}} = \left[\begin{smallmatrix} 0 & \Psi \\ \Psi^* & 0 \end{smallmatrix}\right]$, where $\Psi$ is a unitary operator $\mathcal{U} \to \mathcal{Y}$. The signature operator of the dual supply rate is $J_{\mathrm{imp}*} = \left[\begin{smallmatrix} 0 & \Psi^* \\ \Psi & 0 \end{smallmatrix}\right]$.

(iii) The *transmission* supply rate $j_{\mathrm{tra}}(u, y) = -\langle y, J_{\mathcal{Y}} y\rangle_{\mathcal{Y}} + \langle u, J_{\mathcal{U}} u\rangle_{\mathcal{U}}$ with signature operator $J_{\mathrm{tra}} = \left[\begin{smallmatrix} -J_{\mathcal{Y}} & 0 \\ 0 & J_{\mathcal{U}} \end{smallmatrix}\right]$, where $J_{\mathcal{Y}}$ and $J_{\mathcal{U}}$ are signature operators in $\mathcal{Y}$ and $\mathcal{U}$, respectively. The signature operator of the dual supply rate is $J_{\mathrm{tra}*} = \left[\begin{smallmatrix} -J_{\mathcal{U}} & 0 \\ 0 & J_{\mathcal{Y}} \end{smallmatrix}\right]$.

In the sequel when we talk about *scattering $H$-passive* or *impedance $H$-conservative*, etc., we mean that the supply rate is of the corresponding type. It turns out that although Definition 1.1 and 1.2 can be applied to all three types of supply rates, these three cases still differ significantly from each other.

**2.1. Scattering supply rate.** In the case of scattering supply rate *forward $H$-passivity is equivalent to backward $H$-passivity, hence to passivity*. This is easy to see in the case where $H = 1_{\mathcal{X}}$: the system $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\mathrm{sca}}\right)$ is forward passive if and only if the operator $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is a contraction, which is true if and only if its adjoint $\left[\begin{smallmatrix} A^* & C^* \\ B^* & D^* \end{smallmatrix}\right]$ is a contraction, which is true if and only if the adjoint system $\Sigma^*_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A^* & C^* \\ B^* & D^* \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\mathrm{sca}*}\right)$ is forward passive. The case where $H$ is bounded and has a bounded inverse is almost as easy, and the general case is proved in [AKP05, Proposition 4.6].

The existence of an operator $H > 0$ such that $\Sigma_{\mathrm{i/s/o}}$ is scattering $H$-passive is related to the properties of the transfer function $\Sigma_{\mathrm{i/s/o}}$. To formulate this result we first recall some definitions. The *Schur class* $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ is the unit ball in $H^\infty(\mathcal{U}, \mathcal{Y}, \mathbb{D})$, i.e., each function in $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ is an analytic function on the open unit disk $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ whose values are contractions in $\mathcal{B}(\mathcal{U}, \mathcal{Y})$. The *restricted Schur class* $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \Omega)$, where $\Omega \subset \mathbb{D}$, contains all functions $\theta$ which are restrictions to $\Omega$ of some function in $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$. In other words, $\theta \in \mathcal{S}(\mathcal{U}, \mathcal{Y}; \Omega)$ if the (Nevanlinna–Pick) extension (or interpolation) problem with the (possibly infinite) set of data points $(z, \theta(z))$, $z \in \Omega$, has a solution in $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$. It is known that this problem has a solution if and only if the kernel

$$K^\theta_{\mathrm{sca}}(z, \zeta) = \frac{1_{\mathcal{Y}} - \theta(z)\theta(\zeta)^*}{1 - z\overline{\zeta}}, \quad z, \zeta \in \Omega,$$

is nonnegative definite on $\Omega \times \Omega$, or equivalently, if and only if the kernel

$$K_{\text{sca}}^{\theta *}(z, \zeta) = \frac{1_{\mathcal{U}} - \theta(\zeta)^* \theta(z)}{1 - \overline{\zeta} z}, \quad z, \zeta \in \Omega,$$

is nonnegative definite on $\Omega \times \Omega$ (see [RR82]). We shall here be interested in the case where $\Omega$ is an *open* subset of $\mathbb{D}$, which implies that the solution of this Nevanlinna–Pick extension problem is unique (if it exists).

**Theorem 2.1.** *Let* $\Sigma_{\text{i/s/o}} = \left( \left[ \begin{smallmatrix} A & B \\ C & D \end{smallmatrix} \right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\text{sca}} \right)$ *be an i/s/o system with scattering supply rate and transfer function* $\mathfrak{D}$*, and let* $\Lambda_0(A)$ *be the connected component of* $\Lambda(A) \cap \mathbb{D}$ *which contains the origin.*

(i) *If* $\Sigma_{\text{i/s/o}}$ *is forward* $H$*-passive for some* $H > 0$*, then* $\Sigma_{\text{i/s/o}}$ *is* $H$*-passive and* $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{S}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$.

(ii) *Conversely, if* $\Sigma_{\text{i/s/o}}$ *is minimal and* $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{S}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$*, then* $\Sigma_{\text{i/s/o}}$ *is* $H$*-passive for some* $H > 0$.

In statement (ii) it is actually possible to choose the operator $H$ to satisfy an additional minimality requirement. We shall return to this question in Theorem 3.5.

**2.2. Impedance supply rate.** Also in the case of impedance supply rate *forward* $H$*-passivity is equivalent to backward* $H$*-passivity, hence to passivity*. This is well known in the case where $H = 1_{\mathcal{X}}$ (see, e.g., [Aro79a]). One way to prove this is to reduce the impedance case to the scattering case by means of the following simple transformation.

Suppose that $\Sigma_{\text{i/s/o}} = \left( \left[ \begin{smallmatrix} A & B \\ C & D \end{smallmatrix} \right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\text{imp}} \right)$ is a forward impedance $H$-passive system with signature operator $J_{\text{imp}} = \left[ \begin{smallmatrix} 0 & \Psi \\ \Psi^* & 0 \end{smallmatrix} \right]$. Let $(u, x, y)$ be a trajectory of $\Sigma_{\text{i/s/o}}$. We define a new input $u^\times$ by $u^\times = \frac{1}{\sqrt{2}}(u + \Psi^* y)$ and a new output $y^\times$ by $y^\times = \frac{1}{\sqrt{2}}(\Psi u - y)$, after which we solve (1.2) for $x$ and $y^\times$ in terms of $x_0$ and $u^\times$. It turns out that for this to be possible we need $\Psi + D$ to have a bounded inverse. However, this is always the case, since (1.6) (with $x = 0$) implies that $\Psi^* D + D^* \Psi \geq 0$. A direct computation shows that $(y^\times, x, u^\times)$ is a trajectory of another system $\Sigma_{\text{i/s/o}}^\times = \left( \left[ \begin{smallmatrix} A^\times & B^\times \\ C^\times & D^\times \end{smallmatrix} \right]; \mathcal{U}, \mathcal{X}, \mathcal{Y} \right)$, called the *external Cayley transform of* $\Sigma_{\text{i/s/o}}$, whose coefficients are given by

$$\begin{aligned} A^\times &= A - B(\Psi + D)^{-1} C, & B^\times &= \sqrt{2}\, B(\Psi + D)^{-1} \Psi, \\ C^\times &= -\sqrt{2}\, \Psi(\Psi + D)^{-1} C, & D^\times &= (\Psi - D)(\Psi + D)^{-1} \Psi. \end{aligned} \tag{2.1}$$

The transfer functions of the two systems are connected by

$$\mathfrak{D}^\times(z) = (\Psi - \mathfrak{D}(z))(\Psi + \mathfrak{D}(z))^{-1} \Psi, \quad z \in \Lambda(A) \cap \Lambda(A^\times). \tag{2.2}$$

The external Cayley transform is its own inverse in the sense that $\Psi + D^\times = 2\Psi(\Psi + D)^{-1}\Psi$ always has a bounded inverse, and if we apply the external Cayley transform to the system $\Sigma^\times_{\mathrm{i/s/o}}$, then we recover the original system $\Sigma_{\mathrm{i/s/o}}$.

The main reason for defining the external Cayley transform in the way that we did above is that it 'preserves the energy exchange' in the sense that $j_{\mathrm{imp}}(u, y) = j_{\mathrm{sca}}(y^\times, u^\times)$. This immediately implies that $\Sigma^\times_{\mathrm{i/s/o}}$ is forward scattering $H$-passive whenever $\Sigma_{\mathrm{i/s/o}}$ is forward impedance $H$-passive.[6] According to the discussion in Section 2.1, forward scattering $H$-passivity of $\Sigma^\times_{\mathrm{i/s/o}}$ is equivalent to backward scattering $H$-passivity of $\Sigma^\times_{\mathrm{i/s/o}}$, and this in turn is equivalent to the backward (impedance) $H$-passivity of $\Sigma_{\mathrm{i/s/o}}$. Thus, we get the desired conclusion, namely that forward impedance $H$-passivity implies backward impedance $H$-passivity, hence impedance $H$-passivity.

The same argument can be used to convert all the results mentioned in Section 2.1 into an impedance setting. For simplicity we below take $\mathcal{Y} = \mathcal{U}$ and $\Psi = 1_{\mathcal{U}}$ (this amounts to replacing the output sequence $y$ with values in $\mathcal{Y}$ by the new output sequence $\Psi^* y$ with values in $\mathcal{U}$)). The *Carathéodory class* $\mathcal{C}(\mathcal{U}; \mathbb{D})$ (also called the Carathéodory–Nevanlinna class, or Nevanlinna class, or Weyl class, or Titchmarsh–Weyl class, etc.) consists of all analytic $\mathcal{B}(\mathcal{U})$-valued functions $\psi$ on $\mathbb{D}$ with nonnegative 'real part', i.e., $\psi(z) + \psi(z)^* \geq 0$ for all $z \in \mathbb{D}$. The *restricted Carathéodory class* $\mathcal{C}(\mathcal{U}; \Omega)$, where $\Omega \subset \mathbb{D}$, contains all functions $\theta$ which are restrictions to $\Omega$ of some function in $\mathcal{C}(\mathcal{U}; \mathbb{D})$. In other words, $\theta \in \mathcal{C}(\mathcal{U}; \Omega)$ if the extension problem with the set of data points $(z, \theta(z))$, $z \in \Omega$, has a solution in $\mathcal{C}(\mathcal{U}; \Omega)$. This is equivalent to the requirement that the kernel

$$K^{\psi}_{\mathrm{imp}}(z, \zeta) = \frac{\psi(z) + \psi(\zeta)^*}{1 - z\overline{\zeta}}, \quad z, \zeta \in \Omega,$$

is nonnegative definite on $\Omega \times \Omega$ (this can be proved by reducing the impedance case to the scattering case as explained above).

**Theorem 2.2.** *Let* $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{U}; j_{\mathrm{imp}}\right)$ *be an i/s/o system with impedance supply rate, signature operator* $J_{\mathrm{imp}} = \left[\begin{smallmatrix} 0 & 1_{\mathcal{U}} \\ 1_{\mathcal{U}} & 0 \end{smallmatrix}\right]$, *and transfer function* $\mathfrak{D}$. *Let* $\Lambda_0(A)$ *be the connected component of* $\Lambda(A) \cap \mathbb{D}$ *which contains the origin.*

(i) *If* $\Sigma_{\mathrm{i/s/o}}$ *is forward* $H$*-passive for some* $H > 0$, *then* $\Sigma_{\mathrm{i/s/o}}$ *is* $H$*-passive and* $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{C}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$.

(ii) *Conversely, if* $\Sigma_{\mathrm{i/s/o}}$ *is minimal and* $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{C}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$, *then* $\Sigma_{\mathrm{i/s/o}}$ *is* $H$*-passive for some* $H > 0$.

This theorem follows from Theorem 2.1 as explained above.

[6] It is also true that $\Sigma^\times_{\mathrm{i/s/o}}$ is forward impedance $H$-passive if $\Sigma_{\mathrm{i/s/o}}$ is forward scattering $H$-passive, provided $(\Psi + D)$ has a bounded inverse so that $\Sigma^\times_{\mathrm{i/s/o}}$ exists.

Above we have reduced the impedance passive case to the scattering passive case. Historically the development went in the opposite direction: the impedance version is older than the scattering version. It is related to Neumark's dilation theorem for positive operator-valued measures (see [Bro71, Appendix 1]). In many classical and also in some recent works (especially those where the functions are defined on a half-plane instead of the unit disk) the impedance version is used as 'reference system' from which scattering and other results are derived (see, e.g., [Bro78]). Thus, one easily arrives at the (in my opinion incorrect) conclusion that it does not really matter which one of the two classes is used as the basic corner stone on which the theory is built. However, there is a significant difference between the two classes: the *external Cayley transformation* that converts one of the classes into the other *is well-defined for every impedance $H$-passive system, but not for every scattering $H$-passive system*. In other words, the external Cayley transform maps the class of impedance $H$-passive systems *into but not onto* the class of scattering $H$-passive systems (even if we restrict the input and output dimensions of the scattering system to be the same).

What happens if we try to apply the external Cayley transform to a scattering $H$-passive system for which this transform is not defined (i.e., $\Psi + D$ is not invertible)? In this case the formal transfer function of the resulting system may take its values in the space of closed unbounded operators in $\mathcal{U}$, and it may even be multi-valued. To be able to study this class of 'generalized Carathéodory functions' we need some other more general type of linear systems than the i/s/o systems we have considered so far. This was one of the motivations for the introduction of the notion of a state/signal system in [AS05], to be discussed in Section 3.

**2.3. Transmission supply rate.** In the case of transmission supply rate *forward $H$-passivity is no longer equivalent to backward $H$-passivity*. For simplicity, let us take $H$ to be the identity. Arguing in the same way as in the scattering case we find that $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\mathrm{tra}}\right)$ is forward (transmission) passive if and only if the operator $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is a contraction[7] between two Kreĭn spaces, namely from the space $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{U} \end{smallmatrix}\right]$ with the signature operator $\left[\begin{smallmatrix} 1_{\mathcal{X}} & 0 \\ 0 & J_{\mathcal{U}} \end{smallmatrix}\right]$ to the space $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{Y} \end{smallmatrix}\right]$ with the signature operator $\left[\begin{smallmatrix} 1_{\mathcal{X}} & 0 \\ 0 & J_{\mathcal{Y}} \end{smallmatrix}\right]$. In the same way we find that $\Sigma_{\mathrm{i/s/o}}$ is backward (transmission) passive if $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]^*$ is a contraction from the space $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{Y} \end{smallmatrix}\right]$ with the signature operator $\left[\begin{smallmatrix} 1_{\mathcal{X}} & 0 \\ 0 & J_{\mathcal{Y}} \end{smallmatrix}\right]$ to the space $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{U} \end{smallmatrix}\right]$ with the signature operator $\left[\begin{smallmatrix} 1_{\mathcal{X}} & 0 \\ 0 & J_{\mathcal{U}} \end{smallmatrix}\right]$. However, *in a Kreĭn space setting the contractivity of an operator does not imply that the adjoint of this operator is contractive*, and hence forward transmission passivity does not imply backward transmission passivity without any further restrictions on the system. One necessary condition for the system $\Sigma_{\mathrm{i/s/o}}$ to be both forward and backward (transmission) $H$-passive is that the *dimensions of the negative eigenspaces of $J_{\mathcal{U}}$ and $J_{\mathcal{Y}}$ are the*

[7] An operator $A \in \mathcal{B}(\mathcal{U}; \mathcal{Y})$, where $\mathcal{U}$ and $\mathcal{Y}$ are Kreĭn spaces, is a contraction if $[Au, Au]_{\mathcal{Y}} \le [u, u]_{\mathcal{U}}$ for all $u \in \mathcal{U}$.

*same*. If these dimensions are the same *and finite*, then it is true that forward $H$-passivity is equivalent to backward $H$-passivity, hence to passivity. To prove these statements one can use the following transformation that maps the transmission supply rate into a scattering supply rate.

Suppose that $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\mathrm{tra}}\right)$ is a forward transmission $H$-passive system with signature operator $J_{\mathrm{tra}} = \left[\begin{smallmatrix} J_{\mathcal{Y}} & 0 \\ 0 & J_{\mathcal{Y}} \end{smallmatrix}\right]$. We begin by splitting the output space $\mathcal{Y}$ into the orthogonal direct sum $\mathcal{Y} = -\mathcal{Y}_- \,[\dot{+}]\, \mathcal{Y}_+$, where $\mathcal{Y}_-$ is the negative and $\mathcal{Y}_+$ is the positive eigenspace of $J_{\mathcal{Y}}$. In the same way we split the input space $\mathcal{U}$ into $\mathcal{U} = -\mathcal{U}_- \,[\dot{+}]\, \mathcal{U}_+$, and we split the operator $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ accordingly into

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_1 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right].$$

Let $(u, x, y)$ be a trajectory of $\Sigma_{\mathrm{i/s/o}}$, and split $y$ and $u$ into $y = \left[\begin{smallmatrix} y_- \\ y_+ \end{smallmatrix}\right]$ and $u = \left[\begin{smallmatrix} u_- \\ u_+ \end{smallmatrix}\right]$, so that $y_-$ is a sequence in $\mathcal{Y}_-$, etc. We define a new input $u^{\curvearrowleft}$ by $u^{\curvearrowleft} = \left[\begin{smallmatrix} y_- \\ u_+ \end{smallmatrix}\right]$ and a new output $y^{\curvearrowleft}$ by $y^{\curvearrowleft} = \left[\begin{smallmatrix} u_- \\ y_+ \end{smallmatrix}\right]$, so that $u^{\curvearrowleft}$ is a sequence in $\mathcal{U}^{\curvearrowleft} = \mathcal{Y}_- \oplus \mathcal{U}_+$ and $y^{\curvearrowleft}$ is a sequence in $\mathcal{Y}^{\curvearrowleft} = \mathcal{U}_- \oplus \mathcal{Y}_+$. We then solve (1.2) for $x$ and $y^{\curvearrowleft}$ in terms of $x_0$ and $u^{\curvearrowleft}$. It turns out that for this to be possible we need $D_{11}$ to have a bounded inverse. The forward $H$-passivity of $\Sigma_{\mathrm{i/s/o}}$ implies that $D_{11}$ is injective and has a closed range, but it need not be surjective. However, let us suppose that $D_{11}$ has a bounded inverse. Then a direct computation shows that $(u^{\curvearrowleft}, x, y^{\curvearrowleft})$ is a trajectory of another system $\Sigma^{\curvearrowleft}_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A^{\curvearrowleft} & B^{\curvearrowleft} \\ C^{\curvearrowleft} & D^{\curvearrowleft} \end{smallmatrix}\right]; \mathcal{U}^{\curvearrowleft}, \mathcal{X}, \mathcal{Y}^{\curvearrowleft}\right)$, called the *Potapov–Ginzburg* (or *chain scattering*) *transform of* $\Sigma_{\mathrm{i/s/o}}$, whose coefficients are given by

$$\begin{aligned} \begin{bmatrix} A^{\curvearrowleft} & B^{\curvearrowleft} \\ C^{\curvearrowleft} & D^{\curvearrowleft} \end{bmatrix} &= \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline 0 & 1_{\mathcal{Y}_-} & 0 \\ C_2 & D_{21} & D_{22} \end{array}\right] \left[\begin{array}{c|cc} 1_{\mathcal{X}} & 0 & 0 \\ \hline C_1 & D_{11} & D_{12} \\ 0 & 0 & 1_{\mathcal{U}_+} \end{array}\right]^{-1} \\ &= \left[\begin{array}{c|cc} 1_{\mathcal{X}} & -B_1 & 0 \\ \hline 0 & -D_{11} & 0 \\ 0 & -D_{21} & 1_{\mathcal{Y}_+} \end{array}\right]^{-1} \left[\begin{array}{c|cc} A & 0 & B_2 \\ \hline C_1 & -1_{\mathcal{U}_-} & D_{12} \\ C_2 & 0 & D_{22} \end{array}\right]. \end{aligned} \tag{2.3}$$

The transfer functions of the two systems are connected by

$$\begin{bmatrix} \mathfrak{D}^{\curvearrowleft}_{11}(z) & \mathfrak{D}^{\curvearrowleft}_{12}(z) \\ \mathfrak{D}^{\curvearrowleft}_{21}(z) & \mathfrak{D}^{\curvearrowleft}_{22}(z) \end{bmatrix} = \begin{bmatrix} (\mathfrak{D}_{11}(z))^{-1} & -(\mathfrak{D}_{11}(z))^{-1}\mathfrak{D}_{12}(z) \\ \mathfrak{D}_{21}(z)(\mathfrak{D}_{11}(z))^{-1} & \mathfrak{D}_{22}(z) - \mathfrak{D}_{21}(z)(\mathfrak{D}_{11}(z))^{-1}\mathfrak{D}_{12}(z) \end{bmatrix}, \quad z \in \Lambda(A) \cap \Lambda\big(A^{\curvearrowleft}\big). \tag{2.4}$$

The Potapov–Ginzburg transform is its own inverse in the sense that $D^{\curvearrowleft}_{11} = D_{11}^{-1}$ always has a bounded inverse, and if we apply the Potapov–Ginzburg transform to the system $\Sigma^{\curvearrowleft}_{\mathrm{i/s/o}}$, then we recover the original system $\Sigma_{\mathrm{i/s/o}}$.

The Potapov–Ginzburg transform has been designed to 'preserve the energy exchange' in the sense that $j_{\text{tra}}(u, y) = j_{\text{sca}}(u^{\curvearrowleft}, y^{\curvearrowleft})$. This immediately implies that $\Sigma_{\text{i/s/o}}^{\curvearrowleft}$ is forward scattering $H$-passive whenever $\Sigma_{\text{i/s/o}}$ is forward transmission $H$-passive, provided that $D_{11}$ is invertible so that the transform is defined. As in the impedance case we conclude that the forward transmission $H$-passive system $\Sigma_{\text{i/s/o}}$ is also backward $H$-passive, i.e., $H$-passive, if $D_{11}$ has a bounded inverse (where $D_{11}$ is the part of the feedthrough operator $D$ that maps the negative part of the input space $\mathcal{U}$ into the negative part of the output space $\mathcal{Y}$). The converse is also true: if $\Sigma_{\text{i/s/o}}$ is (transmission) $H$-passive, then $D_{11}$ has a bounded inverse. Thus, a forward transmission $H$-passive system $\Sigma_{\text{i/s/o}}$ is $H$-passive if and only if $D_{11}$ has a bounded inverse, or equivalently, if and only if the Potapov–Ginzburg transform of $\Sigma_{\text{i/s/o}}$ is defined.

The analogue of Theorems 2.1 and 2.2 is more complicated to formulate than in the scattering and impedance cases. In particular, it is not immediately clear how to define the appropriate class of transfer functions. Above we first defined the Schur class $\mathcal{S}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ and the Carathéodory class $\mathcal{C}(\mathcal{U}; \mathbb{D})$ in the full unit disk, and then restricted these classes of functions to some subset $\Omega \subset \mathbb{D}$. Here it is easier to proceed in the opposite direction, and to directly define the *restricted Potapov class* $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \Omega)$ for some $\Omega \subset \mathbb{D}$. We now *interpret $\mathcal{U}$ and $\mathcal{Y}$ as Kreĭn spaces*, i.e., we replace the original Hilbert space inner products in $\mathcal{Y}$ and $\mathcal{U}$ by the Kreĭn space inner products

$$[y, y']_{\mathcal{Y}} = \langle y, J_{\mathcal{Y}} y' \rangle_{\mathcal{Y}}, \quad [u, u']_{\mathcal{U}} = \langle u, J_{\mathcal{U}} u' \rangle_{\mathcal{U}}.$$

In the sequel *we compute all adjoints with respect to these Kreĭn space inner products*, and we also *interpret positivity with respect to these inner products* (so that, e.g., an operator $D$ is nonnegative definite in $\mathcal{U}$ if $[u, Du]_{\mathcal{U}} \geq 0$ for all $u \in \mathcal{U}$). A function $\varphi\colon \Omega \to \mathcal{B}(\mathcal{U}; \mathcal{Y})$ belongs to $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \Omega)$ if both the kernels

$$\begin{aligned} K_{\text{tra}}^{\varphi}(z, \zeta) &= \frac{1_{\mathcal{Y}} - \varphi(z)\varphi(\zeta)^*}{1 - z\overline{\zeta}}, \quad z, \zeta \in \Omega, \\ K_{\text{tra}}^{\varphi *}(z, \zeta) &= \frac{1_{\mathcal{U}} - \varphi^*(\zeta)\varphi(z)}{1 - \overline{\zeta} z}, \quad z, \zeta \in \Omega, \end{aligned} \tag{2.5}$$

are nonnegative definite on $\Omega \times \Omega$.

**Theorem 2.3.** *Let $\Sigma_{\text{i/s/o}} = \left( \left[ \begin{smallmatrix} A & B \\ C & D \end{smallmatrix} \right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}; j_{\text{tra}} \right)$ be an i/s/o system with transmission supply rate, signature operator $J_{\text{tra}} = \left[ \begin{smallmatrix} J_{\mathcal{Y}} & 0 \\ 0 & J_{\mathcal{U}} \end{smallmatrix} \right]$, and transfer function $\mathfrak{D}$. Let $\Lambda_0(A)$ be the connected component of $\Lambda(A) \cap \mathbb{D}$ which contains the origin.*

(i) *If $\Sigma_{\text{i/s/o}}$ is $H$-passive for some $H > 0$, then $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{P}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$.*

(ii) *Conversely, if $\Sigma_{\text{i/s/o}}$ is minimal and $\mathfrak{D}|_{\Lambda_0(A)} \in \mathcal{P}(\mathcal{U}, \mathcal{Y}; \Lambda_0(A))$, then $\Sigma_{\text{i/s/o}}$ is $H$-passive for some $H > 0$.*

This theorem follows from Theorem 2.1 via the Potapov–Ginzburg transformation. Note that (2.5) with $z = \zeta = 0$ implies that both $D$ and $D^*$ are Kreĭn space contractions, so that $D_{11}$ is invertible and the Potapov–Ginzburg transform is defined.

From what we have said so far it seems to follow that the transmission case is not that different from the scattering and impedances cases. However, this impression is not correct. One significant difference is that the Potapov–Ginzburg transformation is not always defined for a forward transmission $H$-passive i/s/o system. Another even more serious problem is that a function in the Potapov class may have singularities inside the unit disk $\mathbb{D}$, which means that in the definition of the (full) Potapov class $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ we must take into account that the function in this class need not be defined everywhere on $\mathbb{D}$. If the negative dimensions of $\mathcal{U}$ and $\mathcal{Y}$ are the same and finite, then this is not a serious problem, because in this case it is possible to define the Potapov class $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ to be the set of all meromorphic functions on $\mathbb{D}$ whose values in $\mathcal{B}(\mathcal{U}; \mathcal{Y})$ are contractive with respect to the Kreĭn space inner products in $\mathcal{U}$ and $\mathcal{Y}$ at all points where the functions are defined. However, in the general case the set of singularities of a function in $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ may be uncountable, and the domain of definition of a function in $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ need not even be connected. For this reason we prefer to define $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ in a different way. We say that a function $\varphi$ belongs to the (full) Potapov class $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$ if it belongs to $\mathcal{P}(\mathcal{U}, \mathcal{Y}; \Omega)$ where *the domain* $\Omega$ *is maximal* in the sense that the function $\varphi$ does not have an extension to any larger domain $\Omega' \subset \mathbb{D}$ with the property that the two kernels in (2.5) are still nonnegative on $\Omega' \times \Omega'$. The existence of such a maximal domain is proved in [AS06b]. This maximal domain need not be connected, but it is still true that if we start from an open set $\Omega \subset \mathbb{D}$, then the values of $\varphi$ on $\Omega$ define the extension of $\varphi$ to its maximal domain uniquely. Moreover, as shown in [AS06b], if $\varphi \in \mathcal{P}(\mathcal{U}, \mathcal{Y}; \mathbb{D})$, then $\varphi$ does not have an analytic extension to any boundary point of its domain contained in the open unit disk $\mathbb{D}$.

Taking a closer look at Theorem 2.3 we observe that it puts one artificial restriction on the transfer function $\mathfrak{D}$, namely that the domain of definition must contain the origin. Not every function in the Potapov class is defined at the origin, so the class of transfer functions covered by Theorem 2.3 is not the full Potapov class. In addition it is possible to extend the Potapov class so that the values of the functions in this class may be unbounded, even multivalued, operators (as in the impedance case) by taking the formal Potapov transforms of functions in $\mathcal{S}(\mathcal{U}, \mathcal{Y}, \mathbb{D})$. Thus, we again see the need of a more general class of systems than the i/s/o class that we have discussed up to now.

## 3. State/signal systems

It is possible to develop a linear systems theory where the differences between the three different types of supply rates, namely scattering, impedance, and transmission, more or less disappear. Both the basic transforms that we have presented above,

namely the external Cayley transform which is used to pass from an impedance $H$-passive system to a scattering $H$-passive system and back, and the Potapov–Ginzburg transform that is used to pass from a transmission $H$-passive system to a scattering $H$-passive system and back, can be regarded as simple 'changes of coordinates in the signal space $\mathcal{W} = \left[\begin{smallmatrix} \mathcal{Y} \\ \mathcal{U} \end{smallmatrix}\right]$'. The main idea is *not to distinguish between the input sequence $u$ and the output sequence $y$*, but to simply regard these as components of the general 'signal sequence' $w = \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$.

We start by combining the input space $\mathcal{U}$ and the output space $\mathcal{Y}$ into one signal space $\mathcal{W} = \left[\begin{smallmatrix} \mathcal{Y} \\ \mathcal{U} \end{smallmatrix}\right]$. This signal space has a natural Kreĭn space[8] inner product obtained from the supply rate $j$ in (1.4), namely

$$\left[\begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} y' \\ u' \end{bmatrix}\right]_{\mathcal{W}} = \left\langle \begin{bmatrix} y \\ u \end{bmatrix}, J \begin{bmatrix} y' \\ u' \end{bmatrix} \right\rangle_{\mathcal{Y}\oplus\mathcal{U}}.$$

If we combine the input sequence $u$ and the output sequence $y$ into one *signal sequence* $w = \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]$, then the basic i/s/o relation (1.1) can be rewritten in the form

$$\begin{bmatrix} x(n+1) \\ x(n) \\ w(n) \end{bmatrix} \in V, \quad n \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\},\ x(0) = x_0, \tag{3.1}$$

where $V$ is the subspace of $\mathfrak{K} := \left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{X} \\ \mathcal{W} \end{smallmatrix}\right]$ given by

$$V = \left\{ \begin{bmatrix} z \\ x \\ w \end{bmatrix} \in \begin{bmatrix} \mathcal{X} \\ \mathcal{X} \\ \mathcal{W} \end{bmatrix} \;\middle|\; \begin{array}{l} z = Ax + Bu, \\ y = Cx + Du, \end{array} \quad w = \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right],\ x \in \mathcal{X},\ u \in \mathcal{U} \right\}. \tag{3.2}$$

It is not difficult to show that the subspace $V$ obtained in this way has the following four properties:

(i) $V$ is closed in $\mathfrak{K}$.

(ii) For every $x \in \mathcal{X}$ there is some $\left[\begin{smallmatrix} z \\ w \end{smallmatrix}\right] \in \left[\begin{smallmatrix} X \\ \mathcal{W} \end{smallmatrix}\right]$ such that $\left[\begin{smallmatrix} z \\ x \\ w \end{smallmatrix}\right] \in V$.

(iii) If $\left[\begin{smallmatrix} z \\ 0 \\ 0 \end{smallmatrix}\right] \in V$, then $z = 0$.

(iv) The set $\left\{ \left[\begin{smallmatrix} x \\ w \end{smallmatrix}\right] \in \left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{W} \end{smallmatrix}\right] \mid \left[\begin{smallmatrix} z \\ x \\ w \end{smallmatrix}\right] \in V \text{ for some } z \in \mathcal{X} \right\}$ is closed in $\left[\begin{smallmatrix} \mathcal{X} \\ \mathcal{W} \end{smallmatrix}\right]$.

**Definition 3.1.** A triple $\Sigma = (V; \mathcal{X}, \mathcal{W})$, where the (*internal*) *state space* $\mathcal{X}$ is a Hilbert space and the (*external*) *signal space* $\mathcal{W}$ is a Kreĭn space and $V$ is a subspace

[8] Both [BS05] and [AS06a] contain short sections on the geometry of a Kreĭn space. For more detailed treatments we refer the reader to one of the books [ADRdS97], [AI89] and [Bog74].

of the product space $\mathfrak{K} := \left[\begin{smallmatrix}\mathcal{Z}\\ \mathcal{X}\\ \mathcal{W}\end{smallmatrix}\right]$ is called a *s/s* (*state/signal*) *node* if it has properties (i)–(iv) listed above. We interpret $\mathfrak{K}$ as a Kreĭn space with the inner product

$$\left[\begin{bmatrix} z\\ x\\ w\end{bmatrix}, \begin{bmatrix} z'\\ x'\\ w'\end{bmatrix}\right]_{\mathfrak{K}} = -\langle z, z'\rangle_{\mathcal{Z}} + \langle x, x'\rangle_{\mathcal{X}} + [w, w']_{\mathcal{W}}, \quad \begin{bmatrix} z\\ x\\ w\end{bmatrix}, \begin{bmatrix} z'\\ x'\\ w'\end{bmatrix} \in \mathfrak{K}, \quad (3.3)$$

and we call $\mathfrak{K}$ the *node space* and $V$ the *generating subspace*.

By a *trajectory* of $\Sigma$ we mean a pair of sequences $(x, w)$ satisfying (3.1). We call $x$ the *state component* and $w$ the *signal component* of this trajectory. By the *s/s system* $\Sigma$ we mean the s/s node $\Sigma$ together with all its trajectories.

The conditions (i)–(iv) above have natural interpretations in terms of the trajectories of $\Sigma$: for each $x_0 \in \mathcal{X}$ condition (ii) gives forward existence of at least one trajectory $(x, w)$ of $\Sigma$ with $x(0) = x_0$. Condition (iii) implies that a trajectory $(x, w)$ is determined uniquely by $x_0$ and $w$, and conditions (i) and (iv) imply that the signal part $x$ depends continuously in $\mathcal{X}^{\mathbb{Z}^+}$ on $x_0 \in \mathcal{X}$ and $w \in \mathcal{W}^{\mathbb{Z}^+}$.

A s/s system $\Sigma$ is *controllable* if the set of all states $x(n)$, $n \geq 1$, which appear in some trajectory $(x, w)$ of $\Sigma$ with $x(0) = 0$ (i.e., an *externally generated trajectory*) is dense in $\mathcal{X}$. The system $\Sigma$ is *observable* if there do not exist any nontrivial trajectories $(x, w)$ where the signal component $w$ is identically zero. Finally, $\Sigma$ is *minimal* if $\Sigma$ is both controllable and observable.

Above we explained how to interpret an i/s/o system $\Sigma_{\mathrm{i/s/o}}$ as a s/s system. Conversely, *from every s/s system* $\Sigma$ *it is possible to create not only one, but infinitely many i/s/o systems*. The representation (3.2) is characterized by the fact that it is a *graph representation of* $V$ *over* $\left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right]$ where $\mathcal{U}$ is one of the two components in a direct sum decomposition of $\mathcal{W} = \mathcal{Y} \dotplus \mathcal{U}$ (not necessarily orthogonal) of $\mathcal{W}$. Indeed, splitting $w$ into $w = \left[\begin{smallmatrix} y\\ u\end{smallmatrix}\right]$ and reordering the components we find that (3.2) is equivalent to

$$V = \left\{ \begin{bmatrix} z\\ y\\ x\\ u\end{bmatrix} \in \begin{bmatrix}\mathcal{Z}\\ \mathcal{Y}\\ \mathcal{X}\\ \mathcal{U}\end{bmatrix} \;\middle|\; \begin{bmatrix} z\\ y\end{bmatrix} = \begin{bmatrix} A & B\\ C & D\end{bmatrix}\begin{bmatrix} x\\ u\end{bmatrix}, \begin{bmatrix} x\\ u\end{bmatrix} \in \begin{bmatrix}\mathcal{X}\\ \mathcal{U}\end{bmatrix} \right\}. \quad (3.4)$$

As shown in [AS05], *the generating subspace of every s/s system* $\Sigma$ *has at least one* (*hence infinitely many*) *graph representation of this type*. A direct sum decomposition $\mathcal{W} = \mathcal{Y} \dotplus \mathcal{U}$ of $\mathcal{W}$ is called an *admissible i/o* (*input/output*) *decomposition of* $\mathcal{W}$ *for* $\Sigma$, or simply an admissible decomposition, if it leads to a graph representation of the generating subspace of $\Sigma$ described above. From each such graph representation of $V$ we get an i/s/o system $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B\\ C & D\end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}\right)$ of $\Sigma$, which we call an *i/s/o representation* of $\Sigma$.

The above definitions are taken from [AS05], [AS06a], and [AS06b]. It turns out that a very large part of the proof of the $H$-passivity theory covered in Section 2 can be carried out directly in the s/s setting, rather than applying the same arguments separately with the scattering, impedance, and transmission supply rates. This leads to both a simplification and to a unification of the whole theory. Below we present the

most basic parts of the $H$-passive s/s theory, and refer the reader to [AS05]–[AS06c] for details.

Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a s/s node. The *adjoint* $\Sigma_* = (V_*; \mathcal{X}, \mathcal{W}_*)$ of $\Sigma$ (introduced in [AS06a, Section 4]) is another s/s node, with the same state space $\mathcal{X}$ as $\Sigma$, and with the signal space $\mathcal{W}_* = -\mathcal{W}$.[9] The generating subspace $V_*$ of $\Sigma_*$ is given by

$$V_* = \left\{ \begin{bmatrix} x_* \\ z_* \\ w_* \end{bmatrix} \,\middle|\, \begin{bmatrix} z_* \\ x_* \\ w_* \end{bmatrix} \in V^{[\perp]} \right\},$$

where $V^{[\perp]}$ is the orthogonal companion to $V$ with respect to the Kreĭn space inner product of $\mathfrak{K}$.[10] The adjoint system $\Sigma_*$ is determined by the property that

$$-\langle x(n+1), x_*(0)\rangle_{\mathcal{X}} + \langle x(0), x_*(n+1)\rangle_{\mathcal{X}} + \sum_{k=0}^{n} [w(k), w_*(n-k)]_{\mathcal{W}} = 0, \quad n \in \mathbb{Z}^+,$$

for all trajectories $(x, w)$ of $\Sigma$.

The following definition is the s/s version of Definitions 1.1 and 1.2.

**Definition 3.2.** Let $H$ be a positive self-adjoint operator in the Hilbert space $\mathcal{X}$. A s/s system $\Sigma$ is

(i) *forward $H$-passive* if $x(n) \in \mathcal{D}(\sqrt{H})$ and

$$\|\sqrt{H}x(n+1)\|_{\mathcal{X}}^2 - \|\sqrt{H}x(n)\|_{\mathcal{X}}^2 \le [w(n), w(n)]_{\mathcal{W}}, \quad n \in \mathbb{Z}^+,$$

for every trajectory $(x, w)$ of $\Sigma$ with $x(0) \in \mathcal{D}(\sqrt{H})$,

(ii) *forward $H$-conservative* if the above inequality holds as an equality,

(iii) *backward $H$-passive* or *$H$-conservative* if $\Sigma_*$ is forward $H^{-1}$-passive or $H^{-1}$-conservative, respectively,

(iv) *$H$-passive* or *$H$-conservative* if it is both forward and backward $H$-passive or $H$-conservative, respectively,

(v) *passive* or *conservative* if it is $1_{\mathcal{X}}$-passive or $1_{\mathcal{X}}$-conservative.

To formulate a s/s version of Theorems 2.1, 2.2 and 2.3 we need a s/s analogue of the transfer function of an i/s/o system. Such an analogue is most easily obtained in the time domain (as opposed to the frequency domain), and it amounts to the introduction of a *behavior*[11] *on the signal space* $\mathcal{W}$. By this we mean a closed right-shift invariant subspace of the Fréchet space $\mathcal{W}^{\mathbb{Z}^+}$. Thus, in particular, the set $\mathfrak{W}$ of all sequences $w$

[9] Algebraically $-\mathcal{W}$ is the same space as $\mathcal{W}$, but the inner product in $-\mathcal{W}$ is obtained from the one in $\mathcal{W}$ by multiplication by the constant factor $-1$.

[10] Thus, $V^{[\perp]} = \{k_* \in \mathfrak{K} \mid [k, k_*]_{\mathfrak{K}} = 0 \text{ for all } k \in V\}$. Note that $V_*$ differs from $V^{[\perp]}$ only by the order of the first two components.

[11] Our behaviors are what Polderman and Willems call *linear time-invariant mainfest behaviors* in [PW98, Definitions 1.3.4, 1.4.1, and 1.4.2]. We refer the reader to this book for further details on behaviors induced by systems with a finite-dimensional state space and for an account of the extensive literature on this subject.

that are the signal parts of externally generated trajectories of a given s/s system $\Sigma$ is a behavior. We call this the *behavior induced by* $\Sigma$, and refer to $\Sigma$ as a *s/s realization of* $\mathfrak{W}$, or, in the case where $\Sigma$ is minimal, as a *minimal s/s realization of* $\mathfrak{W}$. A behavior is *realizable* if it has a s/s realization.

Two s/s systems $\Sigma_1$ and $\Sigma_2$ with the same signal space are *externally equivalent* if they induce the same behavior. This property is related to the notion of *pseudo-similarity*. Two s/s systems $\Sigma = (V; \mathcal{X}, \mathcal{W})$ and $\Sigma_1 = (V_1; \mathcal{X}_1, \mathcal{W})$ are called *pseudo-similar* if there exists an injective densely defined closed linear operator $R\colon \mathcal{X} \to \mathcal{X}_1$ with dense range such that the following conditions hold:

> If $(x(\cdot), w(\cdot))$ is a trajectory of $\Sigma$ on $\mathbb{Z}^+$ with $x(0) \in \mathcal{D}(R)$, then $x(n) \in \mathcal{D}(R)$ for all $n \in \mathbb{Z}^+$ and $(Rx(\cdot), w(\cdot))$ is a trajectory of $\Sigma_1$ on $\mathbb{Z}^+$, and conversely, if $(x_1(\cdot), w(\cdot))$ is a trajectory of $\Sigma_1$ on $\mathbb{Z}^+$ with $x_1(0) \in \mathcal{R}(R)$, then $x_1(n) \in \mathcal{R}(R)$ for all $n \in \mathbb{Z}^+$ and $(R^{-1}x_1(\cdot), w(\cdot))$ is a trajectory of $\Sigma$ on $\mathbb{Z}^+$.

In particular, if $\Sigma_1$ and $\Sigma_2$ are pseudo-similar, then they are externally equivalent. Conversely, if $\Sigma_1$ and $\Sigma_2$ are minimal and externally equivalent, then they are necessarily pseudo-similar. Moreover, a realizable behavior $\mathfrak{W}$ on the signal space $\mathcal{W}$ has a minimal s/s realization, which is determined uniquely by $\mathfrak{W}$ up to pseudo-similarity. (See [AS05, Section 7] for details.)

The *adjoint* of the behavior $\mathfrak{W}$ on $\mathcal{W}$ is a behavior $\mathfrak{W}_*$ on $\mathcal{W}_*$ defined as the set of sequences $w_*$ satisfying

$$\sum_{k=0}^{n} [w(k), w_*(n-k)]_{\mathcal{W}} = 0, \quad n \in \mathbb{Z}^+,$$

for all $w \in \mathfrak{W}$. If $\mathfrak{W}$ is induced by $\Sigma$, then $\mathfrak{W}_*$ is (realizable and) induced by $\Sigma_*$, and the adjoint of $\mathfrak{W}_*$ is the original behavior $\mathfrak{W}$.

The following definition is a s/s analogue of our earlier definitions of the Schur, Carathéodory, and Potapov classes of transfer functions.

**Definition 3.3.** A behavior $\mathfrak{W}$ on $\mathcal{W}$ is

(i) *forward passive* if

$$\sum_{k=0}^{n} [w(k), w(k)]_{\mathcal{W}} \geq 0, \quad w \in \mathfrak{W},\ n \in \mathbb{Z}^+,$$

(ii) *backward passive* if $\mathfrak{W}_*$ is forward passive,

(iii) *passive* if it is realizable[12] and both forward and backward passive.

It is not difficult to see that a s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is forward $H$-passive if and only if $H > 0$ is a solution of the generalized s/s KYP (Kalman–Yakubovich–

[12] We do not know if the realizability assumption is redundant or not.

Popov) inequality[13]

$$\|\sqrt{H}z\|_{\mathcal{X}}^2 - \|\sqrt{H}x\|_{\mathcal{X}}^2 \le [w, w]_{\mathcal{W}}, \quad \begin{bmatrix} z \\ x \\ w \end{bmatrix} \in V, \ x \in \mathcal{D}(\sqrt{H}), \tag{3.5}$$

and that it is forward $H$-conservative if and only if the above inequality holds as an equality.

The following proposition is a s/s version of parts (i) of Theorems 2.1, 2.2, and 2.3.

**Proposition 3.4.** *Let $\mathfrak{W}$ be the behavior induced by a s/s system $\Sigma$.*

(i) *If $\Sigma$ is forward $H$-passive for some $H > 0$, then $\mathfrak{W}$ is forward passive.*

(ii) *If $\Sigma$ is backward $H$-passive for some $H > 0$, then $\mathfrak{W}$ is backward passive.*

(iii) *If $\Sigma$ is forward $H_1$-passive for some $H_1 > 0$ and backward $H_2$-passive for some $H_2 > 0$, then $\Sigma$ is both $H_1$-passive and $H_2$-passive, and $\mathfrak{W}$ is passive.*

The following theorem generalizes parts (ii) of Theorems 2.1, 2.2, and 2.3.

**Theorem 3.5.** *Let $\mathfrak{W}$ be a passive behavior on $\mathcal{W}$. Then*

(i) *$\mathfrak{W}$ has a minimal passive s/s realization.*

(ii) *Every $H$-passive realization $\Sigma$ of $\mathfrak{W}$ is pseudo-similar to a passive realization $\Sigma_H$ with pseudo-similarity operator $\sqrt{H}$. The system $\Sigma_H$ is determined uniquely by $\Sigma$ and $H$.*

(iii) *Every minimal realization of $\mathfrak{W}$ is $H$-passive for some $H > 0$, and it is possible to choose $H$ in such a way that the system $\Sigma_H$ in* (ii) *is minimal.*

Assertion (ii) can be interpreted in the following way: we can always convert an $H$-passive s/s system into a passive one by simply replacing the original norm $\|\cdot\|_{\mathcal{X}}$ in the state space by the new norm $\|x\|_H = \|\sqrt{H}x\|_{\mathcal{X}}$, which is finite for all $x \in \mathcal{D}(\sqrt{H})$, and then completing $\mathcal{D}(\sqrt{H})$ with respect to this new norm.

We shall end this section with a result that says that a suitable subclass of all operators $H > 0$ for which a s/s system $\Sigma$ is $H$-passive can be partially ordered. Here we use the following partial ordering of nonnegative self-adjoint operators on $\mathcal{X}$: if $H_1$ and $H_2$ are two nonnegative self-adjoint operators on the Hilbert space $\mathcal{X}$, then we write $H_1 \preceq H_2$ whenever $\mathcal{D}(H_2^{1/2}) \subset \mathcal{D}(H_1^{1/2})$ and $\|H_1^{1/2}x\| \le \|H_2^{1/2}x\|$ for all $x \in \mathcal{D}(H_2^{1/2})$. For *bounded* nonnegative operators $H_1$ and $H_2$ with $\mathcal{D}(H_2) = \mathcal{D}(H_1) = \mathcal{X}$ this ordering coincides with the standard ordering of bounded self-adjoint operators.

For each s/s system $\Sigma$ we denote the set of operators $H > 0$ for which $\Sigma$ is $H$-passive by $M_\Sigma$, and we let $M_\Sigma^{\min}$ be the set of $H \in M_\Sigma$ for which the system $\Sigma_H$ in assertion (ii) of Theorem 3.5 is minimal.

[13] In particular, in order for the first term in this inequality to be well-defined we require $z \in \mathcal{D}(\sqrt{H})$ whenever $\begin{bmatrix} z \\ x \\ w \end{bmatrix} \in V$ and $x \in \mathcal{D}(\sqrt{H})$.

**Theorem 3.6.** *Let $\Sigma$ be a minimal s/s system with a passive behavior. Then $M_\Sigma^{\min}$ contains a minimal element $H_\circ$ and a maximal element $H_\bullet$, i.e., $H_\circ \preceq H \preceq H_\bullet$ for every $H \in M_\Sigma^{\min}$.*

The two extremal storage functions $E_{H_\circ}$ and $E_{H_\bullet}$ correspond to Willems' [Wil72a], [Wil72b] *available storage* and *required supply*, respectively (there presented in an i/s/o setting). In the terminology of Arov [Aro79b], [Aro95], [Aro99] (likewise in an i/s/o setting), $\Sigma_{H_\circ}$ is the *optimal* and $\Sigma_{H_\bullet}$ is the $*$-*optimal* realization of $\mathfrak{W}$.

## 4. Scattering, impedance and transmission representations of s/s systems

The results presented in Section 2 can be recovered from those in Section 3, together with a number of additional results. This is done by studying different i/s/o representations of a s/s system. Depending on the admissible i/o decomposition of the signal space $\mathcal{W}$ into an input space $\mathcal{U}$ and an output space $\mathcal{Y}$ we get different supply rates (inherited from the Kreĭn space inner product in $\mathcal{W}$).

Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a s/s system, and decompose $\mathcal{W}$ into the direct sum of an input space $\mathcal{U}$ and an output space $\mathcal{Y}$. Furthermore, suppose that this decomposition is admissible, so that it gives rise to an i/s/o representation $\Sigma_{\mathrm{i/s/o}}$ of $\Sigma$. In the case of a *fundamental decomposition* $\mathcal{W} = -\mathcal{Y}\,[\dot{+}]\,\mathcal{U}$, where $\mathcal{Y}$ and $\mathcal{U}$ are Hilbert spaces (i.e., $-\mathcal{Y}$ is an anti-Hilbert space) and $-\mathcal{Y}$ and $\mathcal{U}$ are orthogonal in $\mathcal{W}$, the inner product in $\mathcal{W}$ is given by

$$\left[\begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} y' \\ u' \end{bmatrix}\right]_{\mathcal{W}} = -\langle y, y'\rangle_{\mathcal{Y}} + \langle u, u'\rangle_{\mathcal{U}},$$

which leads to a *scattering supply rate* for the i/s/o representation $\Sigma_{\mathrm{i/s/o}}$. In this case we call $\Sigma_{\mathrm{i/s/o}}$ an *admissible scattering representation* of $\Sigma$. In the case of a (nonorthogonal) *Lagrangian decomposition*, where both $\mathcal{Y}$ and $\mathcal{U}$ are Lagrangian[14] subspaces of $\mathcal{W}$ we get an *impedance supply rate* and an *admissible impedance representation* of $\Sigma$. Finally, if $\mathcal{W} = -\mathcal{Y}\,[\dot{+}]\,\mathcal{U}$ is an arbitrary *orthogonal decomposition* of $\mathcal{W}$ (not necessarily fundamental), then we get a *transmission supply rate* and an *admissible transmission representation* of $\Sigma$. Thus, in the s/s setting the external Cayley transform and the Potapov–Ginzburg transform that we presented in Section 2 are simply two different ways at looking at the same s/s system, via different i/o decompositions of the signal space $\mathcal{W}$ into an input space $\mathcal{U}$ and an output space $\mathcal{Y}$.

The following proposition is related to the discussions at the beginning of Sections 2.1 and 2.2.

**Proposition 4.1.** *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a forward $H$-passive s/s system for some $H > 0$. Then the following claims hold.*

[14] A subspace of a Kreĭn space is Lagrangian if it coincides with its own orthogonal companion.

(i) $\Sigma$ *is $H$-passive if and only if* $\Sigma$ *has an admissible scattering representation, in which case* every *fundamental decomposition of* $\mathcal{W}$ *is admissible.*

(ii) *If* $\Sigma$ *has an admissible impedance representation, then* $\Sigma$ *is $H$-passive.*

The converse of (ii) is not true: there do exist passive s/s systems which do not have any admissible impedance representation, even if we require the positive and negative dimensions of $\mathcal{W}$ to be the same. Every $H$-passive s/s system does have some admissible transmission representations (for example, every scattering representation can be interpreted as a transmission representation), but in general there also exist orthogonal decompositions of the signal space that are not admissible.

One way to prove many of the results listed above is to pass to some particular i/s/o representation $\Sigma_{\mathrm{i/s/o}}$ of the s/s system $\Sigma$, to prove the corresponding result for $\Sigma_{\mathrm{i/s/o}}$, and to reinterpret the result for the s/s system $\Sigma$. In many cases the most convenient choice is to use a scattering representation, corresponding to some admissible fundamental decomposition of the signal space. We recall from Proposition 4.1 that if $\Sigma$ is $H$-passive for some $H > 0$, then every fundamental decomposition is admissible. However, this is not the only possible choice. If $\mathcal{W} = \mathcal{Y} \dot{+} \mathcal{U}$ is an arbitrary admissible i/o decomposition for $\Sigma$, then $\Sigma$ is forward or backward $H$-passive if and only if the corresponding i/s/o system $\Sigma_{\mathrm{i/s/o}}$ is forward or backward $H$-passive with respect to the supply rate on $\mathcal{Y} \dot{+} \mathcal{U}$ inherited from the inner product $[\,\cdot\,,\,\cdot\,]_{\mathcal{W}}$. Thus, in the family of i/s/o systems $\Sigma_{\mathrm{i/s/o}} = \left(\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]; \mathcal{U}, \mathcal{X}, \mathcal{Y}\right)$ that we get from $\Sigma$ by varying the i/o decomposition $\mathcal{W} = \mathcal{Y} \dot{+} \mathcal{U}$ the coefficients $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ vary, and so do the supply rates $j(u, y)$, but *the set of solutions of the generalized KYP inequalities* (1.6) *and* (1.10) *stay the same*.

Up to now we have only considered *admissible* i/o decompositions of the signal space $\mathcal{W}$ of a s/s system $\Sigma$. As we commented earlier, not every Lagrangian or orthogonal decomposition need be admissible for $\Sigma$, even if $\Sigma$ is $H$-passive for some $H > 0$. However, it is still possible to study also these non-admissible decompositions by replacing the i/s/o representations by *left or right affine representations* of $\Sigma$. These are defined for arbitrary decompositions $\mathcal{W} = \mathcal{Y} \dot{+} \mathcal{U}$ (not only for the admissible ones). By a *right affine i/s/o representation of* $\Sigma$ we mean an i/s/o system[15]

$$\Sigma^r_{\mathrm{i/s/o}} = \left(\left[\begin{array}{c|c} A' & B' \\ \hline C'_{\mathcal{Y}} & D'_{\mathcal{Y}} \\ C'_{\mathcal{U}} & D'_{\mathcal{U}} \end{array}\right]; \mathcal{L}, \mathcal{X}, \left[\begin{smallmatrix} \mathcal{Y} \\ \mathcal{U} \end{smallmatrix}\right]\right)$$

with the following two properties: 1) $D' = \left[\begin{smallmatrix} D'_{\mathcal{Y}} \\ D'_{\mathcal{U}} \end{smallmatrix}\right]$ has a bounded left-inverse, and 2) $\left(x, \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]\right)$ is a trajectory of $\Sigma$ if and only if $\left(\ell, x, \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]\right)$ is a trajectory of $\Sigma^r_{\mathrm{i/s/o}}$ for some sequence $\ell$ with values in $\mathcal{L}$. By a *left affine i/s/o representation of* $\Sigma$ we mean

[15] Here the new input space $\mathcal{L}$ is an auxiliary Hilbert space called the *driving variable* space.

an i/s/o system[16]

$$\Sigma^l_{\mathrm{i/s/o}} = \left( \left[\begin{array}{c|cc} A'' & B''_{\mathcal{Y}} & B''_{\mathcal{U}} \\ \hline C'' & D''_{\mathcal{Y}} & D''_{\mathcal{U}} \end{array}\right]; \left[\begin{smallmatrix} \mathcal{Y} \\ \mathcal{U} \end{smallmatrix}\right], \mathcal{X}, \mathcal{K} \right)$$

with the following two properties: 1) $D'' = \left[\begin{array}{cc} \mathfrak{D}''_{\mathcal{Y}} & \mathfrak{D}''_{\mathcal{U}} \end{array}\right]$ has a bounded right-inverse, and 2) $\left(x, \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right]\right)$ is a trajectory of $\Sigma$ if and only if $\left(\left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right], x, 0\right)$ is a trajectory of $\Sigma^l_{\mathrm{i/s/o}}$ (i.e., the output is identically zero in $\mathcal{K}$). The transfer functions of these systems are called the right, respectively left, affine transfer functions of $\Sigma$ corresponding to the i/o decomposition $\mathcal{W} = \mathcal{Y} \dotplus \mathcal{U}$. Note, in particular, that the right and left affine transfer functions are now decomposed into $\mathfrak{D}' = \left[\begin{smallmatrix} \mathfrak{D}'_{\mathcal{Y}} \\ \mathfrak{D}'_{\mathcal{U}} \end{smallmatrix}\right]$ and $\mathfrak{D}'' = \left[\begin{array}{cc} \mathfrak{D}''_{\mathcal{Y}} & \mathfrak{D}''_{\mathcal{U}} \end{array}\right]$, respectively.

Let

$$\Omega(\Sigma^r_{\mathrm{i/s/o}}) = \{z \in \Lambda_{A'} \mid \mathfrak{D}'_{\mathcal{U}}(z) \text{ has a bounded inverse}\},$$
$$\Omega(\Sigma^l_{\mathrm{i/s/o}}) = \{z \in \Lambda_{A''} \mid \mathfrak{D}''_{\mathcal{Y}}(z) \text{ has a bounded inverse}\},$$

and let

$\Omega^r(\Sigma; \mathcal{U}, \mathcal{Y})$ be the union of the above sets $\Omega(\Sigma^r_{\mathrm{i/s/o}})$,

$\Omega^l(\Sigma; \mathcal{U}, \mathcal{Y})$ be the union of the above sets $\Omega(\Sigma^l_{\mathrm{i/s/o}})$.

We can now define the notions of right and left generalized transfer functions of $\Sigma$ with input space $\mathcal{U}$ and output space $\mathcal{Y}$ on the sets $\Omega^r(\Sigma; \mathcal{U}, \mathcal{Y})$ and $\Omega^l(\Sigma; \mathcal{U}, \mathcal{Y})$, respectively, by the formulas

$$\mathfrak{D}_r(z) = \mathfrak{D}'_{\mathcal{Y}}(z)\mathfrak{D}'_{\mathcal{U}}(z)^{-1}, \tag{4.1}$$
$$\mathfrak{D}_l(z) = -\mathfrak{D}''_{\mathcal{Y}}(z)^{-1}\mathfrak{D}''_{\mathcal{U}}(z), \tag{4.2}$$

respectively.

**Theorem 4.2.** *The right-hand side of* (4.1) *does not depend on the choice of* $\Sigma^r_{\mathrm{i/s/o}}$ *as long as* $\Omega(\Sigma^r_{\mathrm{i/s/o}}) \ni z$, *and the right-hand side of* (4.2) *does not depend on the choice of* $\Sigma^l_{\mathrm{i/s/o}}$ *as long as* $\Omega(\Sigma^l_{\mathrm{i/s/o}}) \ni z$.

**Theorem 4.3.** *The right and left generalized transfer functions defined by* (4.1) *and* (4.2), *respectively, coincide on*

$$\Omega(\Sigma; \mathcal{U}, \mathcal{Y}) = \Omega^r(\Sigma; \mathcal{U}, \mathcal{Y}) \cap \Omega^l(\Sigma; \mathcal{U}, \mathcal{Y})$$

(*whenever this set is nonempty*). *If the i/o decomposition* $\mathcal{W} = \mathcal{Y} \dotplus \mathcal{U}$ *is admissible, and if* $A$ *is the main operator of the corresponding i/s/o representation of* $\Sigma$, *then*

$$\Omega^r(\Sigma; \mathcal{U}, \mathcal{Y}) = \Omega^l(\Sigma; \mathcal{U}, \mathcal{Y}) = \Lambda_A,$$

*and the left and right generalized transfer functions coincide with the ordinary transfer function corresponding to the decomposition* $\mathcal{W} = \mathcal{Y} \dotplus \mathcal{U}$

[16]Here the new output space $\mathcal{K}$ is an auxiliary Hilbert space called the *error variable* space.

In the case where the s/s system $\Sigma$ is $H$-passive for some $H > 0$ we can say more. In this case it is possible to choose the different affine representations of $\Sigma$ in such a way that the right and left transfer functions are defined in the whole unit disk $\mathbb{D}$ and belong to $H^\infty$, and they will even be *right and left coprime in* $H^\infty$, respectively. In this way we obtain right and left coprime transmission representations of $\Sigma$, and in the case that the positive and negative dimensions of the signal space $\mathcal{W}$ are the same we also obtain right and left coprime impedance representations. The corresponding right and left coprime affine transfer functions will be generalized Potapov and Carathéodory class functions, respectively.

## 5. Further extensions

The results of Sections 3 and 4 are taken primarily from [AS05], [AS06a]–[AS06c]. At present they do not yet make up a complete theory that would be ready to replace the classical i/s/o theory. However, the following additional *discrete part* ingredients of the s/s theory are presently under active development:

- The study of the *interconnection* of two s/s systems (this is the s/s analogue of feedback).
- *Lossless behaviors* and *bi-lossless extensions* of passive behaviors (including the s/s analogue of Darlington synthesis).
- Additional *representations* of generalized Carathéodory and Potapov class functions.
- External and internal *symmetry* of s/s systems (including reciprocal systems).
- Further studies of the *stability properties* of passive s/s systems.
- Conditions for *ordinary similarity* (as opposed to pseudo-similarity) of minimal passive realizations.

An even larger project is still in its infancy, namely the *extension of the s/s theory to continuous time systems*. Some preliminary results in this direction have been obtained in [BS05] and [MS06a], [MS06b].

## References

[ADRdS97] Alpay, Daniel, Dijksma, Aad, Rovnyak, James, and de Snoo, Henrik, *Schur functions, operator colligations, and reproducing kernel Hilbert spaces*. Oper. Theory Adv. Appl. 96, Birkhäuser, Basel 1997.

[Aro79a] Arov, Damir Z., Optimal and stable passive systems. *Dokl. Akad. Nauk SSSR* **247** (1979), 265–268; English translation in *Soviet Math. Dokl.* **20** (1979), 676–680.

[Aro79b] —, Stable dissipative linear stationary dynamical scattering systems. *J. Operator Theory* **1** (1979), 95–126; English translation in *Interpolation theory, systems theory and related topics*, Oper. Theory Adv. Appl. 134, Birkhäuser, Basel 2002, 99–136.

[Aro95] —, A survey on passive networks and scattering systems which are lossless or have minimal losses. *Archiv für Elektronik und Übertragungstechnik* **49** (1995), 252–265.

[Aro99] —, Passive linear systems and scattering theory. In *Dynamical Systems, Control Coding, Computer Vision* (Padova, 1998), Progr. Systems Control Theory 25, Birkhäuser, Basel 1999, 27–44.

[AKP05] Arov, Damir Z., Kaashoek, Marinus A., and Pik, Derk R., The Kalman–Yakubovich–Popov inequality and infinite dimensional discrete time dissipative systems. *J. Operator Theory* (2005), 46 pages, to appear.

[AS05] Arov, Damir Z., and Staffans, Olof J., State/signal linear time-invariant systems theory. Part I: Discrete time systems. In *The State Space Method, Generalizations and Applications*, Oper. Theory Adv. Appl. 161, Birkhäuser, Basel 2005, 115–177.

[AS06a] —, State/signal linear time-invariant systems theory. Part II: Passive discrete time systems. Submitted, manuscript available at http://www.abo.fi/~staffans/, 2006.

[AS06b] —, State/signal linear time-invariant systems theory. Part III: Transmission and impedance representations of discrete time systems. In preparation, 2006.

[AS06c] —, State/signal linear time-invariant systems theory. Part IV: Affine representations of discrete time systems. In preparation, 2006.

[AI89] Azizov, Tomas Ya., and Iokhvidov, Iosif S., *Linear operators in spaces with an indefinite metric*. Pure Appl. Math. (N. Y.), John Wiley, Chichester 1989.

[BS05] Ball, Joseph A., and Staffans, Olof J., Conservative state-space realizations of dissipative system behaviors. *Integral Equations Operator Theory* **54** (2) (2006), 151–213.

[Bog74] Bognár, János, *Indefinite inner product spaces*. Ergeb. Math. Grenzgeb. 78, Springer-Verlag, Berlin, Heidelberg, New York 1974.

[Bro71] Brodskiĭ, M. S., *Triangular and Jordan representations of linear operators*. Transl. Math. Monogr. 32, Amer. Math. Soc., Providence, RI, 1971.

[Bro78] —, Unitary operator colligations and their characteristic functions. *Russian Math. Surveys* **33** (4) (1978), 159–191.

[IW93] Ionescu, Vlad, and Weiss, Martin, Continuous and discrete-time Riccati theory: a Popov-function approach. *Linear Algebra Appl.* **193** (1993), 173–209.

[Kal63] Kalman, Rudolf E., Lyapunov functions for the problem of Luré in automatic control. *Proc. Nat. Acad. Sci. U.S.A.* **49** (1963), 201–205.

[LR95] Lancaster, Peter, and Rodman, Leiba, *Algebraic Riccati equations*. Oxford Science Publications, The Clarendon Press, Oxford University Press, New York 1995.

[LY76] Lihtarnikov, Andrei L., and Yakubovich, Vladimir A., A frequency theorem for equations of evolution type, *Sibirsk. Mat. Ž.* **17** (5) (1976), 1069–1085, 1198; English translation in *Siberian Math. J.* **17** (1977), 790–803.

[MS06a] Malinen, Jarmo, and Staffans, Olof J., Conservative boundary control systems. Submitted, manuscript available at http://www.abo.fi/~staffans/, 2006.

[MS06b] —, Internal well-posedness of impedance passive boundary control systems. In preparation, 2006.

[Pan99] Pandolfi, Luciano, The Kalman-Yakubovich-Popov theorem for stabilizable hyperbolic boundary control systems. *Integral Equations Operator Theory* **34** (4) (1999), 478–493.

[PAJ91] Petersen, Ian R., Anderson, Brian D. O., and Jonckheere, Edmond A., A first principles solution to the non-singular $H^\infty$ control problem. *Internat. J. Robust Nonlinear Control* **1** (1991), 171–185.

[PW98] Polderman, Jan Willem, and Willems, Jan C., *Introduction to mathematical systems theory: A behavioral approach.* Texts Appl. Math. 26, Springer-Verlag, New York 1998.

[Pop61] Popov, Vasile-Mihai, Absolute stability of nonlinear systems of automatic control. *Avtomat. i Telemeh.* **22** (1961), 961–979; English translation in *Automat. Remote Control* **22** (1961), 857–875.

[RR82] Rosenblum, Marvin, and Rovnyak, James, An operator-theoretic approach to theorems of the Pick–Nevanlinna and Loewner types, II. *Integral Equations Operator Theory* **5** (1982), 870–887.

[Wil72a] Willems, Jan C., Dissipative dynamical systems Part I: General theory. *Arch. Rational Mech. Anal.* **45** (1972), 321–351.

[Wil72b] —, Dissipative dynamical systems Part II: Linear systems with quadratic supply rates. *Arch. Rational Mech. Anal.* **45** (1972), 352–393.

[Yak62] Yakubovich, Vladimir A., The solution of some matrix inequalities encountered in automatic control theory. *Dokl. Akad. Nauk SSSR* **143** (1962), 1304–1307.

[Yak74] —, The frequency theorem for the case in which the state space and the control space are Hilbert spaces, and its application in certain problems in the synthesis of optimal control. I. *Sibirsk. Mat. Ž.* **15** (1974), 639–668, 703; English translation in *Siberian Math. J.* **15** (1974), 457–476 (1975).

[Yak75] —, The frequency theorem for the case in which the state space and the control space are Hilbert spaces, and its application in certain problems in the synthesis of optimal control. II. *Sibirsk. Mat. Ž.* **16** (5) (1975), 1081–1102, 1132; English translation in *Siberian Math. J.* **16** (1974), 828–845 (1976).

Åbo Akademi University, Department of Mathematics, 20500 Åbo, Finland
E-mail: Olof.Staffans@abo.fi
URL: http://www.abo.fi/~staffans/

# Control and numerical approximation of the wave and heat equations

Enrique Zuazua*

*Ama, Magaly, Oihane eta Ainarari.*

**Abstract.** In recent years important progress have been done in the context of numerical approximation of controllability problems for PDEs. It is by now well known that, often, numerical approximation schemes that are stable for solving initial-boundary value problems, develop instabilities when applied to controllability problems. This is due to the presence of spurious high frequency numerical solutions that the control mechanisms are not able to control uniformly as the mesh-size tends to zero. However, the theory is far from being complete. In this article we present some new results in this framework for the wave and the heat equations, which also raise a number of open questions and future directions of research. We first prove that a two-grid method, introduced by R. Glowinski, that is by now well-known to guarantee convergence for the $1-d$ wave equation, also converges in the semilinear setting for globally Lipschitz nonlinearities. This result provides a further evidence of the robustness of the two-grid method. We then show that boundary controls for finite-difference space semi-discretizations of the heat equation converge when applied all along the boundary of the domain, a fact that does not hold for wave-like equations. This confirms that the strong irreversibility of the heat equation enhances the control properties of its numerical approximation schemes. This result fails when the control is restricted to some subsets of the boundary because of the lack of unique continuation of some high frequency eigenvectors of the underlying discrete eigenvalue problem.

**Mathematics Subject Classification (2000).** Primary 93B05; Secondary 35A35, 35K05, 35L05.

**Keywords.** Partial differential equations, finite-difference approximation schemes, controllability, wave equation, heat equation, two-grid method.

## 1. Introduction

In recent years important progresses have been done in the context of numerical approximation of controllability problems for PDEs. It is by now well known that, often, numerical approximation schemes that are stable for solving an initial-boundary value problem, develop instabilities when applied to controllability problems. This is due to the presence of spurious high frequency numerical solutions that the control mechanisms are not able to control uniformly as the mesh-size tends to zero.

*Partially supported by grant MTM2005-00714, the DOMINO Project CIT-370200-2005-10 in the PROFIT program of the MEC (Spain), the SIMUMAT Project S-0505/ESP/0158 of the CAM (Spain) and the European network "Smart Systems".

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

To cure these instabilities a number of methods have been introduced in the literature. We refer to [30] for a recent survey article on the topic.

In this context and in an effort to build a general theory, there are two prototypical equations that need to be understood first of all: the *wave equation* and the *heat equation*.

In the framework of the linear wave equation, R. Glowinski [6] introduced a two-grid control mechanism that allows filtering the high frequency numerical spurious solutions and guarantee the convergence of controls. There are clear numerical evidences of the convergence of the method whose proof has been successfully carried out in [17] in the $1-d$ case by using discrete multipliers. More recently the same result has been proved, with a better estimate on the minimal control time, in [16] by using Ingham type inequalities. Other methods have also been developed for avoiding these instabilities to occur: Tychonoff regularization, Fourier filtering, mixed finite elements,…(see [30]). But most of the existing theory is devoted to linear problems. The first part of this article is devoted to show how the convergence result of the two-grid algorithm can be extended to semilinear systems too, with globally Lipschitz nonlinearities. This result adds one more evidence of the robustness and efficiency of the two-grid algorithm for the control of wave problems.

The high frequency spurious numerical solutions for the wave equation are due to the existence of wave-packets that travel with a vanishing group velocity (see [21], [30]). This can be understood by analyzing the symbol of the operator and the dynamics of the Hamiltonian system generating the bicharacteristic rays. However, one expects that the heat equation, because of its intrinsic time-irreversibility and strong damping should escape to those pathologies and that most common numerical approximation schemes should be controllable, uniformly with respect to the mesh-size. This holds indeed in the $1-d$ setting (see [14]). But, surprisingly enough, this property may fail to hold in $2-d$ even for the simplest finite-difference semi-discretization scheme for the heat equation in the square. This is due to the fact that there are some high-frequency numerical solutions that do not fulfill the classical property of unique continuation of the continuous heat equation. Thus, at the control level, the numerical approximation schemes may generate some solutions which are insensitive to the action of controls. Strictly speaking this happens when the control acts on some (small enough) subsets of the boundary where the equation holds. However, this fact clearly indicates a major difference in the control theoretical behavior of the continuous and the semi-discrete heat equation since the first one is controllable from any open and non-empty subset of the boundary while the second one is not. Characterizing completely the subsets of the boundary for which these pathologies arise is probably a difficult problem. In this article we prove that convergence occurs when the controls act everywhere on the boundary of the domain. This confirms that heat equations are better behaved than wave ones. Indeed, for the wave equation, even if controls act everywhere on the boundary of the domain, the uniform controllability property for numerical approximation schemes may fail because of the existence of spurious numerical solutions that are trapped in the interior of the mesh without reaching the

boundary in an uniform time. Our positive result for the heat equation shows that this kind of spurious solutions are ruled out due to the strong dissipativity of the heat equation and its numerical approximation schemes but, so far, only when the control is distributed everywhere on the boundary.

The lack of unique continuation for semi-discrete heat equations is due to the fact that the property fails to hold for the spectrum of the discrete Laplacian. Indeed, for the Dirichlet spectrum of the continuous Laplacian, unique continuation holds in the sense that, when the normal derivative of an eigenfunction vanishes in a subset of the boundary, the eigenfunction vanishes everywhere. This property fails to hold for the eigenvectors of the discrete Laplacian. The main method to prove unique continuation properties in the continuous framework are the so-called Carleman inequalities. But the discrete analogue is still to be developed. An alternate natural way of addressing this issue, in the spirit of the classical theory of numerical analysis, would consist in viewing the solutions of the discrete problem as a perturbation of those of the continuous one and applying the continuous Carleman inequalities. This approach has been successfully applied in [31] to elliptic equations with irregular coefficients in the principal part. Developing this program in the context of discrete elliptic equations is an interesting open problem.

The two topics we address in this article also raise a number of interesting open problems and future directions of research that we mention briefly. Some of them, in our opinion, are deep and will require important research efforts. The interest of these problems goes much beyond Control Theory since they mainly concern the way classical numerical analysis and the existing theory of partial differential equations have to be melt to address subtle qualitative aspects of numerical solutions. We hope that this article will serve to stimulate research in this area.

## 2. Controllability of the two-grid approximation scheme for the $1-d$ semilinear wave equation

One of the main drawbacks of the existing theory to analyze the controllability of numerical approximation schemes for PDE is that it often relies on Fourier analysis. This makes it of little use for nonlinear problems. However there is by now an extensive literature on the controllability of semilinear PDE and, in particular, of wave and heat equations. Therefore, it is natural to develop numerical methods allowing to address these nonlinear models and to build convergent numerical schemes for their control.

In this section we consider the $1-d$ semilinear wave equation with boundary control:

$$\begin{cases} y_{tt} - y_{xx} + f(y) = 0, & x \in (0,\,1),\ 0 < t < T, \\ y(0,\,t) = 0, \quad y(1,\,t) = v(t), & 0 < t < T, \\ y(x,\,0) = y^0(x),\ y_t(x,\,0) = y^1(x), & x \in (0,\,1). \end{cases} \tag{2.1}$$

Here the control $v = v(t)$ enters into the system through the extreme $x = 1$ of the boundary.

This semilinear wave equation is known to be controllable under sharp growth conditions on the nonlinearity. Namely, if

$$|f'(s)| \leq C \log^2(1 + |s|) \quad \text{for all } s \in \mathbb{R} \tag{2.2}$$

for some $C > 0$, system (2.1) is exactly controllable in any time $T > 2$. This means that for all $(y^0,\ y^1) \in L^2(0,\ 1) \times H^{-1}(0,\ 1)$ and $(z^0,\ z^1) \in L^2(0,\ 1) \times H^{-1}(0,\ 1)$ there exists a control $v \in L^2(0,\ T)$ such that the solution $y$ of (2.1) satisfies

$$y(x,\ T) = z^0(x), \quad y_t(x,\ T) = z^1(x) \quad \text{in } (0,\ 1). \tag{2.3}$$

This result was proved in [26] for $C > 0$ sufficiently small in (2.2) and, without restrictions on the size of the constant $C$, in [1].

This growth condition is sharp since blow-up phenomena may occur for nonlinearities growing faster at infinity and, due to the finite speed of propagation, boundary controls are unable to avoid blow-up to occur. In that case controllability fails.

The most common method to derive the exact controllability property of semilinear equations is based on the following ingredients:

- a fixed point argument;
- sharp estimates on the dependence of controls for the underlying linear equation perturbed by a potential.

We refer to [25] where this method was introduced in the context of the wave equation (see also [22] for further developments, and [23] for an updated survey on this problem) and to [4] where the same technique was applied to semilinear heat equations.

Knowing that the semilinear wave equation (2.1) is controllable under the growth condition (2.2) it is natural to analyze whether the controls can be obtained as limits of controls of numerical approximation schemes. As we have explained in the introduction this issue is delicate even for linear problems, and it is necessarily more complex for nonlinear ones.

Among the possible remedies to the lack of convergence of the standard conservative schemes the two-grid method introduced in [6] seems to be the one that is better adapted to semilinear problems. In this section we confirm this assertion by proving its convergence in this nonlinear setting for globally Lipschitz nonlinearities.

The two-grid scheme is, roughly, as follows.

Given an integer $N \in \mathbb{N}$ we introduce the partition $\{x_j = jh\}_{j=0,\ldots,N+1}$ of the interval $(0,\ 1)$ with $h = 1/(N + 1)$ so that $x_0 = 0$ and $x_{N+1} = 1$.

We then consider the conservative finite-difference semi-discretization of the semilinear wave equation (2.1) as follows:

$$\begin{cases} y_j'' + \frac{2y_j - y_{j+1} - y_{j-1}}{h^2} + f(y_j) = 0, & j = 1, \ldots,\ N,\ 0 < t < T, \\ y_0(t) = 0,\ y_{N+1}(t) = v(t), & 0 < t < T, \\ y_j(0) = y_j^0,\ y_j'(0) = y_j^1, & j = 0, \ldots,\ N + 1. \end{cases} \tag{2.4}$$

The scheme is conservative in the sense that, in the absence of control (i.e. for $v \equiv 0$) the energy of solutions is conserved. The same property holds for the continuous version (2.1). In that case the energy is given by

$$E(t) = \frac{1}{2}\int_0^1 \left[y_t^2(x,t) + y_x^2(x,t)\right] dx + \int_0^1 F(y(x,t))\, dx,$$

where $F$ is a primitive of $f$, i.e. $F(z) = \int_0^z f(s)ds$. In the semi-discrete case the corresponding energy is

$$E_h(t) = \frac{h}{2}\sum_{j=0}^{N}\left[|y_j'|^2 + \left|\frac{y_{j+1}-y_j}{h}\right|^2\right] + h\sum_{j=0}^{N} F(y_j).$$

The goal of this section is to analyze the controllability of (2.4) and whether, as $h \to 0$, the controls of (2.4) converge to those of (2.1). The controls being, in general, non unique, one has to be precise when discussing their convergence. Here, in the linear context, we shall always refer to the controls of minimal $L^2(0,T)$-norm which are given by the so called Hilbert Uniqueness Method (HUM) ([13]). As we mentioned above, in the nonlinear case, the controls we shall deal with are obtained by fixed point methods on the basis of the HUM controls for the linearized problems.

But, even in the linear case, to guarantee convergence as $h \to 0$, the final control requirement has to be relaxed, or the numerical scheme modified.

In [29] it was proved that, if the exact controllability condition is relaxed to the approximate controllability one (in which the state is required to reach an $\varepsilon$-neighborhood of the target), then convergence occurs in the linear framework. But it is convenient to deal with other relaxation criteria that do not introduce extra parameters since the controls may depend on them in a very sensitive way.

The two-grid method is a very natural way of introducing such relaxation. It is based on the idea of relaxing the final condition to avoid the divergence of controls due to the need of controlling high frequency spurious oscillations. To be more precise, the semi-discrete analogue of the exact controllability final condition (2.3) is

$$y_j(T) = z_j^0, \quad y_j'(T) = z_j^1, \quad j = 0, \ldots, N+1. \tag{2.5}$$

But, as it is by now well-known (see [30]), under the final requirement (2.5), controls diverge as $h \to 0$ even for the linear wave equation.

In the two-grid algorithm, the final condition (2.5) is relaxed to

$$\Pi_h\big(Y(T)\big) = \Pi_h(Z^0), \quad \Pi_h\big(Y'(T)\big) = \Pi_h(Z^1), \tag{2.6}$$

where $Y(t)$ and $Y'(t)$ stand for the vector-valued unknowns

$$Y(t) = \big(y_0(t), \ldots, y_{N+1}(t)\big), \quad Y'(t) = \big(y_0'(t), \ldots, y_{N+1}'(t)\big).$$

We shall also use the notation $Y_h$ for $Y$ when passing to the limit to better underline the dependence on the parameter $h$. $\Pi_h$ is the projection operator so that

$$\Pi_h(G) = \left(\frac{1}{2}\left(g_{2j+1} + \frac{1}{2}g_{2j} + \frac{1}{2}g_{2j+2}\right)\right)_{j=0,\dots,\frac{N+1}{2}-1}, \tag{2.7}$$

with $G = (g_0, g_1, \dots, g_N, g_{N+1})$. Note that the projection $\Pi_h(G)$ is a vector of dimension $(N+1)/2$. Thus, roughly speaking, the relaxed final requirement (2.6) only guarantees that half of the state of the numerical scheme is controlled. Despite this fact, the formal limit of (2.6) as $h \to 0$ is still the exact controllability condition (2.3) on the continuous wave equation.

The main result of this section is as follows:

**Theorem 2.1.** *Assume that the nonlinearity $f : \mathbb{R} \to \mathbb{R}$ is such that*

$$f \text{ is globally Lipschitz.} \tag{2.8}$$

*Let $T_0 > 0$ be such that the two-grid algorithm for the control of the linear wave equation converges for all $T > T_0$.*

*Then, the algorithm converges for the semilinear system* (2.1) *too for all $T > T_0$. More precisely, for all $(y^0,\ y^1) \in H^s(0,\ 1) \times H^{s-1}(0,\ 1)$ with $s > 0$, there exists a family of controls $\{v_h\}_{h>0}$ for the semi-discrete system* (2.4) *such that the solutions of* (2.4) *satisfy the relaxed controllability condition* (2.6) *and*

$$v_h(t) \to v(t) \qquad \text{in } L^2(0,\ T),\ h \to 0 \tag{2.9}$$

$$(Y_h,\ Y_h') \to (y,\ y_t) \quad \text{in } L^2(0,\ T;\ L^2(0,\ 1) \times H^{-1}(0,\ 1)) \tag{2.10}$$

*where $y$ is the solution of the semilinear wave equation* (2.1) *and $v$ is a control such that the state $y$ satisfies the final requirement* (2.3).

**Remark 2.2.** Several remarks are in order.

• The controllability of the semilinear wave equation (2.1) under the globally Lipschitz assumption (2.8) on the nonlinearity was proved in [25] in $1-d$ and in the multi-dimensional case. The proof of Theorem 2.1 is based on an adaptation of the arguments in [25] to the two-grid approximation scheme.

Whether the two-grid algorithm applies under the weaker and sharp growth condition (2.2) is an open problem. The difficulty for doing that is that the two existing proofs allowing to deal with the semilinear wave equation under the weaker growth condition (2.2) are based, on a way or another, on the sidewise solvability of the wave equation, a property that the semi-discrete scheme fails to have.

• Theorem 2.1 holds for a sufficiently large time $T_0$. The requirement on $T_0$ is that, in the linear case ($f \equiv 0$), the two-grid algorithm converges for all $T > T_0$. This was proved to hold for $T > 4$ in [17]. The proof in [17] is based on the obtention of the corresponding observability inequality for the solutions of the adjoint semi-discrete wave equation by multiplier techniques. Later on this result was improved in [16]

using a variant of the classical Ingham inequality obtaining the sharp minimal control time $T_0 = 2\sqrt{2}$.

Note that the minimal time for controllability of the continuous wave equation (2.1) is $T = 2$.[1] However this minimal time may not be achieved by the two-grid algorithm as described here since, despite it filters the spurious high frequency numerical solutions, it is compatible with the existence of wave packets travelling with velocity smaller than 1, and this excludes the controllability in the minimal time $T = 2$. The two-grid algorithm can be further improved to get smaller minimal times by considering other projection operators $\Pi_h$, obtained by means of the two-grid approach we shall describe below but with ratio $1/2^\ell$, for some $\ell \geq 2$, instead of the ratio $1/2$. This idea has been used successfully in [7] when proving dispersive estimates for conservative semi-discrete approximation schemes of the Schrödinger equation. When diminishing the ratio between grids, the filtering that the two-grid algorithm introduces concentrates the solutions of the numerical problem on lower and lower frequencies for which the velocity of propagation becomes closer and closer to that of the continuous wave equation. In that way the minimal controllability time may be made arbitrarily close to that of the wave equation $T = 2$ by means of the two-grid approach.

• In the statement of Theorem 2.1 we have chosen initial data for (2.1) in the space $(y^0,\ y^1) \in H^s(0,\ 1) \times H^{s-1}(0,\ 1)$, but we have not explained how the initial data for the semi-discrete system (2.4) have to be taken. The simplest way for doing that is taking as initial data for (2.4) the truncated Fourier series of the continuous initial data $(y^0,\ y^1)$, involving only the first $N$ Fourier modes. One can also define the discrete initial data by taking averages of the continuous ones on the intervals $[x_j - h/2, x_j + h/2]$ around the mesh-points.

• The meaning of the convergence property (2.10) needs also to be made precise. This may be done by extending the semi-discrete state $(Y_h(t),\ Y_h'(t))$ into a continuous one $(y_h(x,\ t),\ y_h'(x,\ t))$ and then proving convergence (2.10) for the extended one. This extension may be defined at least in two different ways. Either by extending the Fourier representation of $Y_h$ or rather by using a standard piecewise linear and continuous extension. We refer to [28] and [19] where these two extensions have been used in similar limit processes.

• In the statement of Theorem 2.1 the initial data are assumed to be in $H^s(0,\ 1) \times H^{-1+s}(0,\ 1)$ for some $s > 0$, which is a slightly stronger regularity assumption than the one needed for the semilinear wave equation (2.1) to be controllable ($L^2(0,\ 1) \times H^{-1}(0,\ 1)$). This is probably a purely technical assumption but it is needed for the method we develop here to apply. The same difficulty arises in the context of the continuous semilinear wave equation [25]. This extra regularity condition for the continuous wave equation was avoided in [1] and [26] but using the very special property of the $1-d$ wave equation of being well-posed in the sidewise sense.

[1] By minimimal control time we mean that the controllability property holds for all time $T$ which is greater than 2. Thus, this does not necessarily mean that controllability occurs for time $T = 2$.

In the context of the problem of numerical approximation we are working this difficulty seems hard to avoid even at the level of passing to the limit as $h \to 0$ on the state equations. Indeed, this requires passing to the limit, in particular, on the nonlinear terms and this seems hard to achieve in the $L^2(0,\ 1) \times H^{-1}(0,\ 1)$-setting because the corresponding states $Y_h$ would then be merely bounded in $C([0,\ T];\ L^2(0,\ 1)) \cap C^1([0,\ T];\ H^{-1}(0,\ 1))$, which seems to be insufficient to guarantee compactness and the convergence of the nonlinear term.

**Proof of Theorem 2.1.** To simplify the presentation we assume that the final target is the null trivial state $z^0 \equiv z^1 \equiv 0$, although the same proof applies in the general case. We proceed in several steps.

*Step* 1. *Two-grid controllability of the semi-discrete system* (2.4). First of all, following the standard fixed point argument ([25]), we prove that, for $h > 0$ fixed, the semilinear system (2.4) is controllable. In fact this argument allows proving that (2.4) is exactly controllable for all $T > 0$. But, as we mentioned above (see [30]), the controls fail to be bounded as $h \to 0$. It is precisely to guarantee that the controls are bounded that we need to relax the final condition to the weaker two-grid one (2.6) and the time $T$ is needed to be large enough as in the statement of Theorem 2.1.

To simplify the presentation we assume that $f \in C^1(\mathbb{R};\ \mathbb{R})$ and $f(0) = 0$, although the proof can be easily adapted to globally Lipschitz nonlinearities. We then introduce the continuous function

$$g(z) = \begin{cases} f(z)/z, & z \neq 0, \\ f'(0), & z = 0. \end{cases} \tag{2.11}$$

Given any semi-discrete function $Z = Z(t) \in C([0,\ T];\ \mathbb{R}^{N+2})$ we consider the linearized wave equation

$$\begin{cases} y_j'' + \frac{2y_j - y_{j+1} - y_{j-1}}{h^2} + g(z_j)y_j = 0, & j = 1, \dots, N;\ 0 < t < T, \\ y_0(t) = y_j^0,\ y_{N+1}(t) = v(t), & 0 < t < T, \\ y_j(0) = y_j'(0) = y_j^1, & j = 0, \dots, N+1. \end{cases} \tag{2.12}$$

We proceed by a classical fixed point argument (see [25]). This requires essentially proving that:

a) For all $Z = Z(t)$ as above (2.12) is two-grid controllable in the sense of (2.6);

b) The mapping $\mathcal{N}(Z) = Y$ has a fixed point.

To be more precise, we shall identify uniquely a control of minimal $L^2(0,\ T)$-norm $v$ (which, obviously, depends on $Z$. Thus, in some cases we shall also denote it as $v_Z$). In this way the controlled trajectory $Y = Y_Z$ will also be uniquely determined and the nonlinear map $\mathcal{N}$ well defined. The problem is then reduced to proving that the

map $\mathcal{N}$ has a fixed point. Indeed, if $Z = Y$, and, consequently, $g(z_j)y_j = f(y_j)$ for all $j = 1, \dots, N$, then $Y$ is also solution of the semilinear semi-discrete equation (2.1) and, of course, satisfies the two-grid relaxed final requirement (2.6).

The existence of the fixed point of $\mathcal{N}$ is consequence of Schauder's fixed point Theorem. The key point to apply it is to show a bound on the two-grid control for the linearized equation (2.12) which is independent of $Z$, i.e. the existence of $C > 0$ such that

$$\|v_Z\|_{L^2(0,T)} \leq C \quad \text{for all } Z \in C([0,\ T];\ \mathbb{R}^{N+2}). \tag{2.13}$$

Here and in the sequel we denote by $v_Z$ the control of the linearized system (2.12) to underline the fact that the control depends on the potential $g(Z)$ and thus on $Z$.

To do that we argue as in [17], reducing the problem to the obtention of a suitable observability inequality for the adjoint system:

$$\begin{cases} \varphi_j'' + \frac{2\varphi_j - \varphi_{j+1} - \varphi_{j-1}}{h^2} + g(z_j)\varphi_j = 0, & j = 1, \dots, N,\ 0 < t < T, \\ \varphi_0 = \varphi_{N+1} = 0, & 0 < t < T, \\ \varphi_j(T) = \varphi_j^0,\ \varphi_j'(T) = \varphi_j^1, & j = 1, \dots, N. \end{cases} \tag{2.14}$$

For doing that, however, system (2.14) has to be considered only in the class of slowly oscillating data obtained as extensions to the fine grid (the original one, of size $h$) of data defined on a coarse grid of size $2h$. In other words, we consider the class of data

$$\mathcal{V}_h = \left\{ \Phi = (\varphi_0, \dots, \varphi_{N+1}) : \varphi_{2j+1} = \frac{\varphi_{2j} + \varphi_{2j+2}}{2},\ j = 0, \dots, \frac{N-1}{2} \right\}. \tag{2.15}$$

Note that any vector in $\mathcal{V}_h$ is completely determined by its values on the grid of mesh-size $2h$. Implicitly we are assuming that $1/2h$ is an integer number so that $(N-1)/2 = 1/2h - 1$ is an integer too.

In [17] and [16] it was proved that for $T > T_0$, where $T_0$ is as in Remark 2.2, the following observability inequality holds:

$$E_0 \leq C \int_0^T \left| \frac{\varphi_N}{h} \right|^2 dt, \tag{2.16}$$

with $C > 0$ independent of $h > 0$ and for all solution $\Phi = (\varphi_0, \dots, \varphi_{N+1})$ of (2.14) with data $(\Phi^0,\ \Phi^1) \in \mathcal{V}_h \times \mathcal{V}_h$ when $g \equiv 0$. Here $E_0$ stands for the total energy of solutions at time $t = T$, which is constant in time when $g \equiv 0$:

$$E(t) = \frac{h}{2} \sum_{j=0}^{N} \left[ |\varphi_j'|^2 + \left| \frac{\varphi_{j+1} - \varphi_j}{h} \right|^2 \right]. \tag{2.17}$$

At this point it is important to emphasize that the key ingredient of the proof of convergence for the two-grid algorithm for the linear wave equation is precisely that the observability constant $C$ in (2.16) is uniform, independent of $h$.

Let us now address the perturbed problem (2.14). We first observe that, because of the globally Lipschitz assumption on $f$, the function $g$ is uniformly bounded, i.e.

$$\|g\|_{L^\infty(\mathbb{R})} \leq L, \tag{2.18}$$

$L$ being the Lipschitz constant of $f$. Therefore for all $Z \in C([0,\ T];\ \mathbb{R}^{N+2})$ it follows that

$$\|g(Z)\|_{L^\infty(0,\ T;\ \mathbb{R}^{N+2})} \leq L. \tag{2.19}$$

System (2.14) can then be viewed as a family of perturbed semi-discrete wave equations, the perturbations with respect to the conservative wave equation being a family of zero order bounded potentials. Then a standard perturbation argument allows showing that (2.16) holds for system (2.14) too, with, possibly, a larger observability constant, depending on $L$, but independent of $Z$. In fact, arguing by contradiction, since $h > 0$ is fixed and we are therefore dealing with finite-dimensional dynamical systems, the problem is reduced to show that the following unique continuation property holds for all $Z$:

$$\text{If } \varphi_N(t) = 0,\ 0 < t < T,\ \text{ then } \Phi \equiv 0. \tag{2.20}$$

This property is easy to prove by induction. Indeed, using the boundary condition $\varphi_{N+1} \equiv 0$ and the fact that $\varphi_N \equiv 0$, and writing the equation (2.14) for $j = N$ we deduce that

$$\frac{\varphi_{N-1}}{h^2} = \varphi_N'' + \frac{2\varphi_N - \varphi_{N-1}}{h^2} + g(z_N)\varphi_N = 0. \tag{2.21}$$

This implies that $\varphi_{N-1} = 0$. Repeating this argument we deduce that $\Phi \equiv 0$.

Once (2.16) holds for the solutions of (2.14) with initial data in $\mathcal{V}_h \times \mathcal{V}_h$, uniformly on $Z$, system (2.14) turns out to be controllable in the sense of (2.6) with an uniform bound on the control, independent of $Z$, i.e.

$$\|v\|_{L^2(0,\ T)} \leq C(h,\ \|(Y^0,\ Y^1)\|,\ L,\ T) \quad \text{for all } Z. \tag{2.22}$$

We emphasize that the bound (2.22), in principle, depends on the time of control $T$, the mesh-size $h$, the Lipschitz constant $L$ of the nonlinearity $f$ and the norm of the initial data to be controlled, but is independent of $Z$.

As a consequence of (2.22) a similar estimate can be obtained for the state $Y$ solution of (2.12), i.e.

$$\|Y\|_{C^1([0,\ T];\ \mathbb{R}^{N+2})} \leq C \quad \text{for all } Z. \tag{2.23}$$

This allows applying the Schauder's fixed point theorem to the map $\mathcal{N}$, which turns out also to be compact from $L^2(0,\ T;\ \mathbb{R}^{N+2})$ into itself, thanks to (2.23). In this way we conclude that, for all $h > 0$, system (2.4) is controllable in the sense of (2.6).

*Step* 2. *Uniform controllability with respect to h.* In the previous step we have proved the controllability of (2.4) but with estimates on controls and states depending on $h$.

In order to pass to the limit as $h \to 0$ we need to get a bound on controls and states which is independent of $h$. For doing that we need to assume that $T > T_0$ (so that the linear unperturbed numerical schemes are uniformly, with respect to $h > 0$, two-grid controllable) and that the initial data $(y^0,\ y^1)$ belong to $H^s(0,\ 1) \times H^{s-1}(0,\ 1)$.

The last requirement is important to get the compactness of the nonlinear term. Indeed, in that setting the control for the continuous wave equation (2.1) belongs to $H^s(0,\ T)$ rather than $L^2(0, T)$ and the controlled trajectory $y$ then belongs to $C([0,\ T];\ H^s(0,\ 1)) \cap C^1([0,\ T];\ H^{s-1}(0,\ 1))$. This guarantees the required compactness properties to deal with the nonlinear term $f(y)$ in (2.1). Indeed, when passing to the limit, the pointwise convergence of the state in $(0,\ 1) \times (0,\ T)$ is needed and this is achieved by means of the extra $H^s$ regularity imposed on the initial data (see [25]). This is necessary both when treating the continuous equation (2.1) by fixed point arguments and also when dealing with numerical approximation issues and limit processes as $h \to 0$.

To analyze the controllability of the systems under consideration in $H^s(0,\ 1) \times H^{s-1}(0,\ 1)$, we first need to analyze the $H^{-s}$-version of the observability inequality (2.16), namely:

$$E_{0,\ -s} \leq C_s \left\| \frac{\varphi_N}{h} \right\|^2_{H^{-s}(0,\ T)}. \tag{2.24}$$

Inequality (2.24) may be proved for the adjoint system (2.14) in the absence of the potential induced by the nonlinearity, i.e. for

$$\begin{cases} \psi_j'' + \frac{2\psi_j - \psi_{j+1} - \psi_{j-1}}{h^2} = 0, & j = 1, \dots, N,\ 0 < t < T, \\ \psi_0 = \psi_{N+1} = 0, & 0 < t < T, \\ \psi_j(T) = \psi_j^0,\ \psi_j(T) = \psi_j^1, & j = 1, \dots, N. \end{cases} \tag{2.25}$$

More precisely, for $0 < s < 1/2$ and $T > T_0$ as in Theorem 2.1, there exists a constant $C_s$ such that (2.24) holds for all solution $\psi$ of (2.25) with initial data in $\mathcal{V}_h \times \mathcal{V}_h$ and all $h > 0$. We emphasize that the constant $C_s$ is independent of $h$.

In (2.24) $E_{0,\ -s}$ stands for the $H^{-s}$ version of the energy of system (2.25), which is constant in time. It can be defined easily by means of the Fourier expansion of solutions and it is then the discrete analogue of the continuous energy

$$E_{0,\ -s} = \frac{1}{2}\left[ \left\| \psi^0 \right\|^2_{H^{1-s}(0,\ 1)} + \left\| \psi^1 \right\|^2_{H^{-s}(0,\ 1)} \right] \tag{2.26}$$

which is constant in time for the solutions of the unperturbed adjoint wave equation

$$\begin{cases} \psi'' - \psi_{xx} = 0, & 0 < x < 1, 0 < t < T, \\ \psi(0,\ t) = \psi(1,\ t) = 0, & 0 < t < T, \\ \psi(x,\ T) = \psi^0(x),\ \psi_t(x,\ T) = \psi^1(x), & 0 < x < 1. \end{cases} \tag{2.27}$$

The inequality (2.24) may be obtained, as (2.16), by the two methods mentioned above:

- It can be proved as a consequence of (2.16) directly using interpolation arguments (see, for instance, [25]);
- It can also be obtained by the variant of the Ingham inequality in [16].

Once (2.24) is proved for the unperturbed system (2.25), uniformly on $h > 0$, we are in conditions to prove it for the perturbed system (2.14) uniformly on $h > 0$ and $Z$ too. To do it we use a classical perturbation and compactness argument (see [25]).

We decompose the solution $\Phi$ of (2.14) as $\Phi = \Psi + \Sigma$ where $\Psi$ solves the unperturbed system (2.25) with the same data $(\Phi^0,\ \Phi^1)$ as $\Phi$ itself and where the remainder $\Sigma = (\sigma_0, \dots, \sigma_{N+1})$ solves

$$\begin{cases} \sigma_j'' + \frac{2\sigma_j - \sigma_{j+1} - \sigma_{j-1}}{h^2} = -g(z_j)\varphi_j, & j = 1, \dots, N,\ 0 < t < T, \\ \sigma_0 = \sigma_{N+1} = 0, & 0 < t < T, \\ \sigma_j(T) = \sigma_j'(T) = 0, & j = 1, \dots, N. \end{cases} \tag{2.28}$$

In view of (2.24), which is valid for $\Psi$, we deduce that

$$E_{0,\,-s} \leq 2C_s \Big[ \Big\| \frac{\varphi_N}{h} \Big\|^2_{H^{-s}(0,\,T)} + \Big\| \frac{\sigma_N}{h} \Big\|^2_{H^{-s}(0,\,T)} \Big]. \tag{2.29}$$

Using discrete multipliers (see [8]) it follows that

$$\Big\| \frac{\sigma_N}{h} \Big\|_{L^2(0,\,T)} \leq C \| \{ g(z_j)\varphi_j \} \|_{L^2(0,\,T;\,\ell^2_h)} \tag{2.30}$$

with a constant $C$ which depends on $T$ but is independent of $h$.

In (2.30) we use the notation

$$\| \{p_j\} \|_{L^2(0,\,T;\,\ell^2_h)} = \Big[ h \int_0^T \sum_{j=1}^N p_j^2(t)\, dt \Big]^{1/2} \tag{2.31}$$

which is simply the $L^2$-norm, scaled to the mesh-size $h > 0$.

Combining (2.30)–(2.31) and using that the nonlinearity $g$ is uniformly bounded we deduce that

$$E_{0,\,-s} \leq C \Big[ \Big\| \frac{\varphi_N}{h} \Big\|^2_{H^{-s}(0,\,T)} + \| \Phi \|^2_{L^2(0,\,T;\,\ell^2_h)} \Big], \tag{2.32}$$

for every solution $\Phi$ of (2.14) with data $(\Phi^0,\ \Phi^1)$ in $\mathcal{V}_h \times \mathcal{V}_h$, every $h > 0$ and $Z$.

To conclude we can apply a compactness-uniqueness argument whose details may be found in [28] where it was fully developed in the context of the $2-d$ semi-discrete wave equation. It consists simply in showing, by contradiction, that there exists an uniform constant $C > 0$ such that

$$\| \Phi \|_{L^2(0,\,T;\,\ell^2_h)} \leq C \Big\| \frac{\varphi_N}{h} \Big\|_{H^{-s}(0,\,T)} \tag{2.33}$$

for every solution $\Phi$ of (2.14), every $h > 0$ and $Z$. To do it we assume that there exists a sequence $h \to 0$, potentials of the form $g(Z_h)$ and initial data in $\mathcal{V}_h \times \mathcal{V}_h$ for which (2.33) fails and, consequently,

$$\left\| \frac{\varphi_N}{h} \right\|_{H^{-s}(0,\,T)} \to 0, \quad h \to 0 \tag{2.34}$$

$$\|\Phi\|_{L^2(0,\,T;\,\ell^2_h)} = 1. \tag{2.35}$$

Combining (2.32), (2.34) and (2.35), the corresponding sequence of data $(\Phi^0_h,\ \Phi^1_h)$ turns out to be bounded in $H^{1-s}(0,\,1) \times H^{-s}(0,\,1)$ (at this point we are implicitly working with the piecewise linear extension of the data). By the well-posedness of (2.14) in these spaces the corresponding solutions $\Phi_h$ turn out to be bounded in $L^\infty(0,\ T;\ H^{s-1}(0,\,1)) \cap W^{1,\,\infty}(0,\ T;\, H^{-s}(0,\,1))$. Therefore, they are relatively compact in $L^2(0,\ T;\ L^2(0,\,1))$. Passing to the limit as $h \to 0$ we obtain a solution $\varphi$ of an adjoint wave equation of the form

$$\begin{cases} \varphi'' - \varphi_{xx} + a(x,\,t)\varphi = 0, & 0 < x < 1,\ 0 < t < T, \\ \varphi(0,\,t) = \varphi(1,\,t) = 0, & 0 < t < T, \\ \varphi(x,\,T) = \varphi^0(x),\ \varphi_t(x,\,0) = \varphi^1(x), & 0 < x < 1, \end{cases} \tag{2.36}$$

such that

$$\partial_x \varphi(1,\,t) = 0, \quad 0 < t < T \tag{2.37}$$

and

$$\|\varphi\|_{L^2(0,\,T;\,L^2(0,\,1))} = 1. \tag{2.38}$$

This is clearly a contradiction since, in view of the unique continuation property of the solutions of the wave equation (2.36), (2.37) implies that $\varphi \equiv 0$, which is incompatible with (2.38). This is true because $T > T_0$ and, in particular $T > 2$.

In this argument the bounded limit potential $a = a(x,\,t)$ in (2.36) arises as weak-$*$ limit of the discrete ones $g\,(Z_h)$, (of its piecewise linear extension to $0 < x < 1,\ 0 < t < T$, to be more precise). Therefore $a$ also fulfills the bound $\|a\|_\infty \leq L$, $L$ being the Lipschitz constant of $f$.

For this argument to apply one needs to pass to the limit in the potential perturbation $g\,(Z_h)\,\Phi_h$ in (2.14). This can be done because of the strong convergence of $\Phi_h$ (of its extension to $0 < x < 1$) in $L^2((0,\ 1) \times (0,\ T))$.

Once (2.24) is known to hold for all $h > 0$ and all data in $\mathcal{V}_h \times \mathcal{V}_h$ this allows proving the uniform controllability of (2.4) in the spaces $H^s(0,\,1) \times H^{-1+s}(0,\,1)$, in the two-grid sense (2.6). This can be done applying the fixed point argument in Step 1. More precisely, it follows that there exists a family of controls $v_h \in H^s(0,\ T)$, with an uniform bound

$$\|v_h\|_{H^s(0,\,T)} \leq C \left\| \left(Y^0_h,\ Y^1_h\right) \right\|_{H^s(0,1) \times H^{s-1}(0,1)} \tag{2.39}$$

such that the solutions $Y_h$ of (2.4) satisfy (2.6).

*Step* 3. *Two-grid observability* $\Longrightarrow$ *Two-grid controllability.* For the sake of completeness, let us show how the two-grid control of (2.12) can be obtained as a consequence of the observability inequality (2.24) for the solutions of the adjoint wave equation (2.14) with initial data in the class $\mathcal{V}_h \times \mathcal{V}_h$ in (2.15) of slowly oscillating data.

We introduce the functional

$$J_h\left(\Phi^0,\ \Phi^1\right) = \frac{1}{2}\left\| \frac{\varphi_N}{h} \right\|^2_{H^{-s}(0,\,T)} + h\sum_{j=1}^{N}\left[y_j^0\varphi_j'(0) - y_j^1\varphi_j(0)\right], \tag{2.40}$$

which is continuous and convex. Moreover, in view of (2.24), the functional $J_h\colon \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{R}$ is uniformly coercive. Let us denote by $(\Phi_h^{0,*},\ \Phi_h^{1,*})$ the minimizer of $J_h$ over $\mathcal{V}_h \times \mathcal{V}_h$. Then,

$$\left\langle DJ_h\left(\Phi_h^{0,*},\ \Phi_h^{1,*}\right),\ \left(\Phi^0,\ \Phi^1\right)\right\rangle = 0 \tag{2.41}$$

for all $(\Phi^0,\ \Phi^1) \in \mathcal{V}_h \times \mathcal{V}_h$. This implies that

$$\left(\frac{\varphi_N^*}{h},\ \frac{\varphi_N}{h}\right)_{H^{-s}(0,\,T)} + h\sum_{j=1}^{N}\left[y_j^0\varphi_j'(0) - y_j^1\varphi_j(0)\right] = 0 \tag{2.42}$$

where $\Phi_h^*$ stands for the solution of (2.14) with the minimizer $(\Phi_h^{0,*},\ \Phi_h^{1,*})$ as data and $\Phi$ the solution with data $(\Phi^0,\ \Phi^1)$.

We now choose the control

$$v_h = I_s\frac{\varphi_N^*}{h}, \tag{2.43}$$

where $I_s\colon H^{-s}(0,\ T) \to H^s(0,\ T)$ is the canonical duality isomorphism.

Equation (2.42) then reads

$$\int_0^T v_h\frac{\varphi_N}{h}dt + h\sum_{j=0}^{N}\left[y_j^0\varphi_j'(0) - y_j^1\varphi_j(0)\right] = 0. \tag{2.44}$$

On the other hand, using (2.43) as control in (2.12), multiplying by $\Phi$ the solution of the adjoint system (2.14), adding on $j = 1, \ldots,\ N$ and integrating by parts with respect to $t \in (0,\ T)$, we deduce that

$$\int_0^T v\frac{\varphi_N}{h}dt + h\sum_{j=1}^{N}\left[y_j^0\varphi_j'(0) - y_j^1\varphi_j(0)\right] - h\sum_{j=1}^{N}\left[y_j(T)\varphi_j^1 - y_j'(T)\varphi_j^0\right] = 0. \tag{2.45}$$

Combining (2.44)–(2.45) we deduce that the solution $Y_h$ of (2.12) satisfies

$$h\sum_{j=1}^{N}\left[y_j(T)\varphi_j^1 - y_j'(T)\varphi_j^0\right] = 0 \quad \text{for all } (\varphi^0,\ \varphi^1) \in \mathcal{V}_h \times \mathcal{V}_h. \tag{2.46}$$

This means that both $Y_h(T)$ and $Y_h'(T)$ are perpendicular to $\mathcal{V}_h$. This is equivalent to the two-grid control requirement (2.6) with $z^0 \equiv z^1 \equiv 0$.

In view of this construction and using the observability inequality (2.24), which is uniform with respect to $h > 0$ and $Z$, we can obtain uniform bounds on the controls. Indeed, by (2.43) we have

$$\|v_h\|_{H^s(0,\,T)} = \left\| \frac{\varphi_N^*}{h} \right\|_{H^{-s}(0,\,T)}. \tag{2.47}$$

On the other hand, the minimizer $(\Phi_h^{0,*},\ \Phi_h^{1,*})$ is such that

$$J\left(\Phi_h^{0,*},\ \Phi_h^{1,*}\right) \leq 0 \tag{2.48}$$

and this implies

$$\begin{aligned} \frac{1}{2}\|\frac{\varphi_N^*}{h}\|^2_{H^{-s}(0,\,T)} &\leq \left| h \sum_{j=1}^{N} \left[ y_j^0 \varphi_j^{*'}(0) - y_j^1 \varphi_j^*(0) \right] \right| \\ &\leq \|(Y^0,\ Y^1)\|_{H^s(0,1)\times H^{s-1}(0,1)} \sqrt{E^*_{0,\,-s}} \end{aligned} \tag{2.49}$$

where $E^*_{0,\,-s}$ denotes the $E_{0,\,-s}$ energy of the minimizer $(\Phi_h^{0,*},\ \Phi_h^{1,*})$.

Combining (2.48), (2.49) and the observability inequality (2.24) we deduce that

$$\|v_h\|_{H^s(0,\,T)} \leq 2\sqrt{C_s}\|(Y^0,\ Y^1)\|_{H^s(0,1)\times H^{s-1}(0,1)}, \tag{2.50}$$

where $C_s$ is the same constant as in (2.24). In particular, the bound (2.50) on the control is independent of $h > 0$ and $Z$.

*Step* 4. *Passing to the limit.* Using the uniform bound (2.39) it is easy to pass to the limit and get the null-controllability of the semilinear wave equation (2.1).

Indeed, as a consequence of (2.39) and by the well-posedness of (2.4) we deduce that the controlled state $Y_h$ is uniformly bounded in $L^\infty(0,\ T;\ H^s(0,\ 1)) \cap W^{1,\infty}(0,\ T;\ H^{s-1}(0,\ 1))$.

By extracting subsequences we have

$$v_h \rightharpoonup v \text{ weakly in } H^s(0,\ T) \tag{2.51}$$

$$Y_h \rightharpoonup y \text{ weakly in } L^2(0,\ T;\ H^s(0,\ 1)) \cap H^1(0,\ T;\ H^{s-1}(0,\ 1)). \tag{2.52}$$

Consequently, in particular,

$$v_h \to v \text{ strongly in } L^2(0,\ T) \tag{2.53}$$

$$Y_h \to y \text{ strongly in } L^2((0,\ 1) \times (0,\ T)). \tag{2.54}$$

These convergences suffice to pass to the limit in (2.4) and to get (2.1). The strong convergence (2.54) is particularly relevant when doing that since it allows passing to the limit in the nonlinearity.

Here the convergence of the states $Y_h$ may be understood in the sense that its extensions to functions defined for all $0 < x < 1$ converge.

One can also check that the limit state $y = y(x, t)$ satisfies the final exact controllability requirement (2.3) as a consequence of the two-grid relaxed version (2.6) that the semi-discrete state $Y_h$ satisfies. This can be done either by transposition or by compactness.

This concludes the sketch of the proof of Theorem 2.1.

**Remark 2.3.** Several remarks are in order:

- The proof we have given can be adapted to other equations and schemes. In particular it applies to the two-grid finite element approximation of (2.1).
- In [2] a mixed finite-element discretization scheme has been introduced for which the uniform controllability property holds without requiring any filtering or two-grid adaptation. The arguments we have developed here can also be adapted to prove convergence of that method in the semilinear case under the globally Lipschitz assumption on the nonlinearity $f$.
- In [12] it was proved that the standard finite-difference semi-discretization for the exact controllability of the following beam equation converges without filtering or two-grid adaptation:

  $$y_{tt} + y_{xxxx} = 0.$$

  The method of proof of Theorem 2.1 allows showing that the same is true in the semilinear context too.
- The proof of convergence of the two-grid control algorithm is still to be developed for numerical approximations of the wave equation in the multi-dimensional case. But, in view of the proof of Theorem 2.1, which can be easily adapted to the multi-dimensional framework, we can say that, if convergence is proved in the linear case, the same will hold in the semilinear one too, for globally Lipschitz nonlinearities.
- For the semilinear wave equation (2.1) the *local* null-controllability can be proved in wider classes of nonlinearities satisfying $f'(0) = 0$. Here by local null-controllability we refer to the property that sufficiently small initial data can be driven to the null state, i.e. to the existence of $\delta > 0$ such that the control driving the solution to the final equilibrium $\{0, 0\}$ exists for all initial data $\{y^0, y^1\}$ such that

  $$\|y^0\|_{L^2(0,1)} + \|y^1\|_{H^{-1}(0,1)} \leq \delta.$$

  It can be proved as a consequence of the controllability of the linear wave equation applying the inverse function theorem around the null state. In order to guarantee the well-posedness of the semilinear wave equation (2.1) in

$L^2(0,\,1) \times H^{-1}(0,\,1)$ one also needs to impose a growth condition on the nonlinearity of the form

$$|f'(s)| \leq C|s| \quad \text{for all } s \in \mathbb{R}.$$

But this allows proving local controllability for quadratic nonlinearities, for instance (see [24]).

The method of proof of Theorem 2.1 can be used to prove the convergence of the two-grid method in that context of local controllability too.

- In [24] it was also observed that for nonlinearities with the good sign property, for instance, for
$$f(s) = |s|^{p-1}s \quad \text{for all } s \in \mathbb{R}$$
with $1 < p \leq 2$, every initial datum may be driven to zero for a sufficiently large time. For doing that one first uses the exponential decay of solutions with boundary feedback ([9]) to later apply the local controllability property when the solution becomes small enough.

  To adapt that result to the framework of the two-grid scheme one could need a uniform (with respect to $h$) stabilization result for the numerical schemes with boundary feedback. However it is well known that, due to the lack of uniform boundary observability as $h \to 0$, the uniform stabilization property fails. In [19] and [20] (see also [18]) the uniform (with respect to $h$) exponential decay property was proved but by adding a viscous damping term distributed all along the mesh. In view of the efficiency of the two-grid approach at the level of controllability, one would expect the uniform (with respect to $h > 0$) exponential decay property to hold for initial data in the space $\mathcal{V}_h \times \mathcal{V}_h$ without the extra viscous damping term. But this property does not seem to happen. Indeed, when trying to obtain the uniform exponential decay from the uniform observability inequality a technical difficulty appears since the space $\mathcal{V}_h \times \mathcal{V}_h$ is not invariant under the flow of the semi-discrete wave equation. Thus, the observability inequality we have obtained in the time interval $[0,\,T]$ with $T > T_0$ can not be extended for all $t \geq 0$, a fact that would be needed for proving the exponential decay. But in fact, the situation is much worse and the uniform exponential decay fails to hold since, despite of the fact that the two-grid initial data have a distribution of energy so that most of it is concentrated on the low frequencies, as time evolves, this partition of energy is lost because, precisely, high frequency components are weakly dissipated. The apparently purely technical difficulty for proving the uniform decay turns out to be in fact the reason for the lack of uniform decay.

- Recently it has also been proved that $1 - d$ semilinear wave equations are controllable in the sense that two different equilibria can be connected by a controlled trajectory provided they belong to the same connected component of

the set of stationary solutions (see [3]). This holds without any sign restriction on the nonlinearity and therefore without excluding blow-up phenomena to occur. Proving the convergence of the two-grid algorithm in what concerns that result is an open problem too.

- The main drawback of the arguments we have used in the proof of Theorem 2.1 is that they do not provide any explicit estimate on the cost of controlling the system with respect to the Lipschitz constant $L$ of the nonlinearity. This is due to the use of compactness-uniqueness arguments in the obtention of the uniform (with respect to $h$) observability estimates (2.24) in the class of initial data $\mathcal{V}_h \times \mathcal{V}_h$. Therefore we may not recover by this method the property of controllability of the semilinear wave equation under the sharp superlinear growth condition (2.2).

- The existence of the convergent controls of the semi-discrete semilinear system has been proved by means of a fixed point method. This adds extra technical difficulties for its efficient computation. The most common tool to deal with such problem, for $h$ fixed and after a suitable time discretization, is the Newton method with variable step. It is applicable in the present situation since the nonlinear map $\mathcal{N}$ under consideration is differentiable when the nonlinearity $f$ in the equation is $C^1$. In each iteration of the Newton method, one is lead to solve a linearized control problem. But this one is solvable by means of a standard conjugate gradient algorithm ([6]) because of the uniform observability properties that are guaranteed to hold, as we have seen, due to the two-grid relaxation we have introduced. A complete numerical study of these issues is yet to be developed.

- The limit control $v$ we have obtained can be proved to be a fixed point of the nonlinear map $\mathcal{N}$ that corresponds to the controllability of the semilinear continuous wave equation, based on the HUM controls of minimal $L^2(0, T)$-norm for the linearized wave equations. Thus, the controls we are dealing with both, for the continuous and the semi-discrete equation, belong to the same category.

## 3. Boundary control of the finite-difference space semi-discretizations of the heat equation

**3.1. Problem formulation.** This section is devoted to analyze the null controllability of the space semi-discretizations, by means of finite differences, of the heat equation in multi-dimensional domains. To simplify the presentation we focus on the $2 - d$ case although the same results, with similar proofs, apply in any dimension $d \geq 2$.

The heat equation in bounded domains is known to be null-controllable from any open, non-empty subset of the domain or its boundary [5]. Thus, it is natural to

analyze whether the control is the limit of the controls of the semi-discrete systems as the mesh-size tends to zero. But this turns out not to be the case even for the heat equation in the square when the control is applied on a strict subset of one of the segments constituting its boundary (see [30]).

In this section we prove a positive counterpart of that result. More precisely, we prove that convergence holds when the control acts on a whole side of the boundary. The proof uses the Fourier series development of solutions, which allows reducing the problem to a one-parameter family of controllable $1-d$ heat equations. As a consequence of that result we can prove convergence for general domains when the control is applied on the whole boundary. For that it is sufficient to extend the initial data in the original domain to data in a square containing it and then obtaining the controls on the boundary of the original domain as restrictions to the boundary of the states defined in the extended square.

The same results hold in any space dimension.

To be more precise, let $\Omega$ be the square $\Omega = (0,\pi)\times(0,\pi)$ of $\mathbb{R}^2$. Let $\Gamma_0$ be one side of its boundary, say $\Gamma_0 = \{(x_1,0) : 0 < x_1 < \pi\}$.

Consider the heat equation with control on $\Gamma_0$:

$$\begin{cases} y_t - \Delta y = 0 & \text{in } \Omega\times(0,T), \\ y = 0 & \text{on } [\partial\Omega\setminus\Gamma_0]\times(0,T), \\ y = v & \text{on } \Gamma_0\times(0,T), \\ y(x,0) = y^0(x) & \text{in } \Omega. \end{cases} \tag{3.1}$$

Here $y = y(x,t)$, with $x = (x_1,x_2)$, is the *state* and $v = v(x_1,t)$ is the *control*.

System (3.1) is well-known to be null-controllable in any time $T > 0$ (see Fursikov and Imanuvilov [5] and Lebeau and Robbiano [11]). More precisely, the following holds: *For any $T > 0$ and any $y^0 \in L^2(\Omega)$ there exists $v \in L^2(\Gamma_0\times(0,T))$ such that the solution $y = y(x,t)$ of* (3.1) *satisfies*

$$y(x,T) \equiv 0. \tag{3.2}$$

*Moreover, there exists a constant $C > 0$ depending on $T$ but independent of the initial datum $y^0$ such that*

$$\|v\|_{L^2(\Gamma_0\times(0,T))} \le C\|y^0\|_{L^2(\Omega)} \quad \text{for all } y^0 \in L^2(\Omega). \tag{3.3}$$

In fact the same result holds in a general smooth bounded domain $\Omega$ and with controls in any open non-empty subset $\Gamma_0$ of its boundary.

In the present setting, this result is equivalent to an observability inequality for the *adjoint heat equation*:

$$\begin{cases} \varphi_t + \Delta\varphi = 0 & \text{in } \Omega\times(0,T), \\ \varphi = 0 & \text{on } \partial\Omega\times(0,T), \\ \varphi(x,T) = \varphi^0(x) & \text{in } \Omega. \end{cases} \tag{3.4}$$

More precisely, it is equivalent to the existence of a positive constant $C > 0$ such that

$$\|\varphi(0)\|^2_{L^2(\Omega)} \leq C \int_0^T \int_{\Gamma_0} \left| \frac{\partial \varphi}{\partial n} \right|^2 d\sigma dt \quad \text{for all } \varphi^0 \in L^2(\Omega). \tag{3.5}$$

Here and in the sequel by $n$ we denote the unit exterior normal vector field and by $\partial \cdot / \partial n$ the normal derivative. In this case, over $\Gamma_0$, $\partial \cdot / \partial n = -\partial \cdot / \partial_{x_2}$.

Let us now consider the finite-difference space semi-discretizations of (3.1) and (3.4).

Given $N \in \mathbb{N}$ we set $h = \pi/(N+1)$ and we consider the mesh

$$x_{i,j} = (ih, jh), \quad i, j = 0, \ldots, N+1. \tag{3.6}$$

We now introduce the finite-difference semi-discretizations:

$$\begin{cases} y'_{j,k} + \frac{1}{h^2}(4y_{j,k} - y_{j+1,k} - y_{j-1,k} - y_{j,k+1} - y_{j,k-1}) = 0, \ (j,k) \in \Omega_h, \ 0 < t < T, \\ y_{j,k} = 0, \ (j,k) \in [\partial\Omega \setminus \Gamma_0]_h, \quad 0 < t < T, \\ y_{j,0} = v_j, \ j = 0, \ldots, N+1, \quad 0 < t < T, \\ y_{j,k}(0) = y^0_{j,k}, \qquad\qquad (j,k) \in \Omega_h, \end{cases} \tag{3.7}$$

and

$$\begin{cases} \varphi'_{j,k} - \frac{1}{h^2}(4\varphi_{j,k} - \varphi_{j+1,k} - \varphi_{j-1,k} - \varphi_{j,k+1} - \varphi_{j,k-1}) = 0, \ (j,k) \in \Omega_h, \ 0 < t < T, \\ \varphi_{j,k} = 0, \qquad (j,k) \in [\partial\Omega]_h, \ 0 < t < T, \\ \varphi_{j,k}(T) = \varphi^0_{j,k}, \quad (j,k) \in \Omega_h. \end{cases} \tag{3.8}$$

To simplify the notations, we have denoted by $\Omega_h$ (resp. $\partial\Omega_h$) the set of interior (resp. boundary) nodes, and by $[\partial\Omega \setminus \Gamma_0]_h$ the set of indices $(j,k)$ so that the corresponding nodes belong to $\partial\Omega \setminus \Gamma_0$. Here and in the sequel $y_{j,k} = y_{j,k}(t)$ (resp. $\varphi_{j,k} = \varphi_{j,k}(t)$) stands for an approximation of the solution $y$ of (3.1) (resp. $\varphi$ of (3.4)) at the mesh-points $x_{i,j}$. On the other hand, $v_j$ denotes the control that acts on the semi-discrete system (3.7) through the subset $[\Gamma_0]_h$ of the boundary. Note that the control does not depend of the index $k$ since the subset of the boundary $[\Gamma_0]_h$ where the control is being applied corresponds to $k = 0$.

In order to simplify the notation we introduce the vector unknowns and control

$$Y_h = (y_{j,k})_{0 \leq j,k \leq N+1}, \quad \Phi_h = (\phi_{j,k})_{0 \leq j,k \leq N+1}, \quad V_h = (v_j)_{1 \leq j \leq N}, \tag{3.9}$$

that we shall often denote simply by $Y$, $\Phi$ and $V$.

Accordingly, systems (3.7) and (3.8) read as follows:

$$\begin{cases} Y'_h + A_h Y_h = B_h V_h, \\ Y_h(0) = Y^0_h, \end{cases} \tag{3.10}$$

$$\begin{cases} \Phi_h' - A_h \Phi_h = 0, \\ \Phi_h(T) = \Phi_h^0. \end{cases} \tag{3.11}$$

We denote by $A_h$ the usual positive-definite symmetric matrix associated with the five-point finite-different scheme we have employed in the discretization of the Laplacian so that

$$(A_h W)_{j,k} = \frac{1}{h^2}(4w_{j,k} - w_{j+1,k} - w_{j-1,k} - w_{j,k+1} - w_{j,k-1}), \tag{3.12}$$

for the inner nodes. In (3.11) the homogenous boundary conditions have been integrated by assuming simply that their values in the expression (3.12) have been replaced by the zero one. On the other hand the linear operator $B_h$ in (3.10) is such that the action of the control $v_j$ enters on those nodes which are neighbors to those of $[\Gamma_0]_h$, i.e. for $k = 1$, so that $[B_h V]_{j,k} = 0$ whenever $2 \leq k \leq N$ but $[B_h V]_{j,1} = -v_j/h^2$.

The null-controllability problem for system (3.10) reads as follows: Given $Y^0 \in \mathbb{R}^{N+2\times N+2}$ to find $V \in L^2(0, T;\ \mathbb{R}^N)$ such that the solution $Y$ of (3.10) satisfies

$$Y(T) = 0. \tag{3.13}$$

On the other hand, the problem of observability for system (3.11) consists in proving the existence of $C > 0$ such that

$$\|\Phi(0)\|_h^2 \leq Ch \int_0^T \sum_{j=1}^{N} \left| \frac{\phi_{j,1}}{h} \right|^2 dt \tag{3.14}$$

for every solution $\Phi$ of (3.11).

In (3.14) $\|\cdot\|_h$ stands for the scaled Euclidean norm

$$\|\Phi\|_h = \left[ h^2 \sum_{j,k=0}^{N+1} |\phi_{j,k}|^2 \right]^{1/2} \tag{3.15}$$

and the right hand side term of inequality (3.14) represents the discrete version of the $L^2$-norm of the normal derivative in (3.5).

A similar problem can be formulated in general bounded smooth domains $\Omega$. In that case, obviously, the domain $\Omega$ needs to be approximated by domains $\Omega_h$ whose boundaries are constituted by mesh-points. We first address the case of the square domain by Fourier series to later derive some consequences for general domains.

All this section is devoted to the problem of null control. Obviously the situation is different if the final requirement is relaxed to an approximate controllability condition. In that context, as a consequence of the null controllability of the limit heat equation and the convergence of the numerical algorithm it can be proved that the state $Y_h$ at time $t = T$ can be driven to a final state of norm $\varepsilon_h$ such that $\varepsilon_h \to 0$ as $h \to 0$. But, as mentioned above, this property fails in general in the framework of null controllability. At this point the work in [10] is also worth mentioning. There it was proved that, in the context of analytic semigroups, one can also get uniform bounds on the number of iterations needed for computing controls using conjugate gradient algorithms.

**3.2. The square domain.** The goal of this subsection is to prove that, as $h \to 0$, the controls $V_h$ of (3.10) are uniformly bounded and converge in $L^2(\Gamma_0 \times (0, T))$ to the control of (3.1). All along this section we deal with controls of minimal $L^2$-norm, the so-called HUM controls.

In order to make this convergence result more precise it is convenient to take the following facts into account:

- To state and analyze the convergence of the discrete states $Y_h$ it is convenient to extend them to continuous functions $y_h(x, t)$ with respect to the space variable $x = (x_1, x_2)$. This can be done, as in the previous section, in two different ways either by considering a piecewise linear and continuous extension or extending the discrete Fourier expansion of solutions by keeping exactly the same analytic expression. The control $V_h$ has to be extended as well to a function depending on the continuous variable $0 < x_1 < \pi$. This can be done in the same two ways.

- To state the convergence of controls as $h \to 0$ the initial data $Y_h^0$ in (3.10) have to be chosen in connection with the initial data $y^0$ of the PDE (3.1). This may be done in several ways. When $y^0$ is continuous, $Y_h^0$ can be taken to be the restriction of $y^0$ to the mesh-points. Otherwise, one can take average values over cells, or simply truncate the Fourier expansion of the continuous initial datum $y^0$ by taking the first $N \times N$ terms.

This being made precise, the following result holds:

**Theorem 3.1.** *Let $T > 0$ be any positive control time. Let $y^0 \in L^2(\Omega)$ and $Y_h^0$ be as above. Then, the null controls $V_h$ for the semi-discrete problem* (3.10) *are uniformly bounded, with respect to h and converge in $L^2(\Gamma_0 \times (0, T))$ towards the null control of the heat equation* (3.1). *The semi-discrete controlled states $Y_h$ also converge to the controlled state y of the heat equation in $L^2(0, T; H^{-1}(\Omega))$ satisfying the null final condition* (3.2).

**Remark 3.2.** The result is sharp in what concerns the support $\Gamma_0$ of the control. Indeed, as pointed out in [30] this result fails when $[\Gamma_0]_h$ is replaced by the set of indices $[\Gamma_0^*]_h$ in which the first node corresponding to the index $j = 1$ is removed. In that case the observability inequality (3.14) fails because of the existence of a non-trivial solution (3.11) such that $\Phi$ vanishes on $[\Gamma_0^*]_h$. This is so in fact because of the existence of a non-trivial eigenvector of the discrete Laplacian $A_h$ with eigenvalue $\lambda_h = 4/h^2$, taking alternating values $\pm 1$ along the diagonal and vanishing out of it.

The main elements of the proof of this result are the following. The key point is precisely proving that the observability inequality (3.14) is uniform with respect to the mesh-size $h > 0$. Once this is done standard variational methods allow proving that the controls are uniformly bounded and then passing to the limit as $h \to 0$. We refer to [27] where the same issue was addressed for the heat equation in thin cylindrical

domains by similar tools and to [14] where the limit process was described in detail in the context of the finite-difference semi-discrete approximation of the $1-d$ heat equation.

The method of proof of the uniform estimate (3.14) depends heavily on the Fourier decomposition of solutions. To develop it we need some basic facts about the Fourier decomposition of the discrete Laplacian.

The eigenvalue problem associated with the semi-discrete system (3.11) is as follows:

$$\begin{cases} \frac{1}{h^2}\left[4w_{j,k} - w_{j+1,k} - w_{j-1,k} - w_{j,k+1} - w_{j,k-1}\right] = \lambda w_{j,k}, & (j,k) \in \Omega_h, \\ w_{j,k} = 0, & (j,k) \in [\partial\Omega]_h. \end{cases} \tag{3.16}$$

Its spectrum may be computed explicitly:

$$\lambda^{\ell,m}(h) = \frac{4}{h^2}\left[\sin^2\left(\frac{\ell h}{2}\right) + \sin^2\left(\frac{mh}{2}\right)\right] \tag{3.17}$$

$$W^{\ell,m}(h) = w^{\ell,m}(x)\big|_{x=(jh,kh),\, j,k=0,\cdots,N+1} \tag{3.18}$$

for $\ell, m = 1, \ldots, N$, where $w^{\ell,m}(x)$ are the eigenfunctions of the continuous Laplacian:

$$w^{\ell,m}(x) = \frac{2}{\pi}\sin(\ell x_1)\sin(m x_2).$$

In particular, in view of (3.18) the eigenvectors of the discrete system (3.16) are simply the restrictions of the eigenfunctions of the continuous Laplacian to the mesh points. Of course, this is a very particular fact that is not true for general domains $\Omega$.

It is also easy to check that

$$\lambda^{\ell,m}(h) \to \lambda^{\ell,m} = \ell^2 + m^2 \quad \text{as } h \to 0 \tag{3.19}$$

for all $\ell, m \geq 1$, where $\lambda^{\ell,m}$ stand for the eigenvalues of the continuous Laplacian. This confirms that the 5-point finite-difference scheme provides a convergent numerical scheme.

The eigenvectors $\{W^{\ell,m}\}_{\ell,m=1,\ldots,N}$ constitute an orthonormal basis of $\mathbb{R}^{N\times N}$ with respect to the scalar product

$$\langle f, \widetilde{f}\rangle_h = \left[h^2 \sum_{j,k=1}^{N} f_{j,k}\widetilde{f}_{j,k}\right]^{1/2}, \tag{3.20}$$

associated with the norm (3.15).

The solution of the semi-discrete adjoint system (3.11) can also be easily developed in this basis:

$$\Phi_h(t) = \sum_{\ell,m=1}^{N} a^{\ell,m} e^{-\lambda^{\ell,m}(h)(T-t)} W^{\ell,m} \tag{3.21}$$

where $\{a^{\ell,m}\}$ are the Fourier coefficients of the datum at time $t = T$:

$$\Phi_h^0 = \sum_{\ell,m=1}^{N} a^{\ell,m} W^{\ell,m}, \quad a^{\ell,m} = \langle \Phi_h^0,\ W^{\ell,m} \rangle_h. \tag{3.22}$$

Solutions may also be rewritten in the form

$$\Phi_h(t) = \sum_{\ell=1}^{N} \psi^m(t) \otimes \sigma^m, \tag{3.23}$$

where

$$\sigma^m = \left( \frac{\sqrt{2}}{\sqrt{\pi}} \sin(mkh) \right)_{k=0,\dots,N+1},$$

so that $W^{\ell,m} = \sigma^\ell \otimes \sigma^m$, and each vector-valued function $\psi^m(t) = (\psi_j^m(t))_{j=0,\dots,N+1}$ is a solution of the $1-d$ semi-discrete problem:

$$\begin{cases} \psi_j' - \left[2\psi_j - \psi_{j+1} - \psi_{j-1}\right]/h^2 + \mu^m \psi_j = 0, & j = 1, \dots, N,\ 0 < t < T, \\ \psi_0 = \psi_{N+1} = 0, & 0 < t < T, \\ \psi_j(T) = \psi_j^0, & j = 1, \dots, N, \end{cases} \tag{3.24}$$

where $\mu^m = \frac{4}{h^2} \sin^2\left(\frac{mh}{2}\right)$.

The observability inequality (3.14) is equivalent to proving the $1-d$ analogue for (3.24), uniformly with respect to the index $m \geq 1$, i.e.

$$\|\psi(0)\|_h^2 \leq C \int_0^T \left| \frac{\psi_1}{h} \right|^2 dt, \tag{3.25}$$

for all $\psi^0$, $\psi$ being the solution of (3.24), with a constant $C > 0$ which is independent of $m$.

The proof of this $1-d$ uniform estimate can be developed easily following the arguments in [14]. In fact that inequality is an immediate consequence of the explicit form of the spectrum together with a technical result on series of real exponentials that we recall for the sake of completeness. Consider the class $\mathcal{L}(\xi, M)$ constituted by increasing sequences of positive real numbers $\{\nu_j\}_{j\geq 1}$ such that

$$\nu_{j+1} - \nu_j \geq \xi > 0 \quad \text{for all } j \geq 1, \tag{3.26}$$

$$\sum_{k \geq M(\delta)} \frac{1}{\nu_k} \leq \delta \qquad \text{for all } \delta > 0. \tag{3.27}$$

Here $\xi$ is any positive number and $M\colon (0, \infty) \to \mathbb{N}$ is a function such that $M(\delta) \to \infty$ as $\delta \to 0$. Obviously, different values of $\xi$ and $M$ determine different classes of sequences $\mathcal{L}(\xi, M)$. The following holds (see [14]):

**Proposition 3.3.** *Given a class of sequences $\mathcal{L}(\xi, M)$ and $T > 0$, there exists a positive constant $C > 0$ such that*

$$\int_0^T \Big| \sum_{k=1}^{\infty} a_k e^{-\nu_k t} \Big|^2 dt \geq \frac{C}{\left[\sum_{k\geq 1} 1/\nu_k\right]} \sum_{k\geq 1} \frac{|a_k|^2 e^{-2\nu_k T}}{\nu_k}, \tag{3.28}$$

*for all $\{\nu_k\}_{k\geq 1} \in \mathcal{L}(\xi, M)$ and all bounded sequence $\{a_k\}_{k\geq 1}$.*

Note that the sequences of eigenvalues of problems (3.24) belong to the same class $\mathcal{L}(\xi, M)$ for all $h > 0$ and $m \geq 1$. Thus, the constant $C$ in (3.28) is uniform and, consequently, (3.24) holds, with an observability constant independent of $h > 0$ and $m \geq 1$ as well.

**Remark 3.4.** The same result holds for the case in which the control acts as a right hand side external force applied on a band, i.e. on a set of the form $\omega = \{(x_1, x_2) : 0 < x_1 < \gamma, 0 < x_2 < \pi\}$ with $0 < \gamma < \pi$. The corresponding continuous model reads

$$\begin{cases} y_t - \Delta y = f 1_\omega & \text{in } \Omega \times (0, T), \\ y = 0 & \text{on } \partial\Omega \times (0, T), \\ y(x, 0) = y^0(x) & \text{in } \Omega, \end{cases} \tag{3.29}$$

where $f = f(x_1, x_2, t)$ is the control and $1_\omega$ is the characteristic function of the set $\omega$ where the control is applied.

The corresponding observability inequality is

$$\|\varphi(0)\|^2_{L^2(\Omega)} \leq C \int_0^T \int_\omega \varphi^2 \, dx dt \quad \text{for all } \varphi^0 \in L^2(\Omega). \tag{3.30}$$

The problems can be formulated similarly for the semi-discrete scheme we have considered.

The observability inequality (3.30) and the corresponding semi-discrete versions hold uniformly with respect to the mesh-size parameter $h > 0$. Consequently the heat equation (3.29) and the corresponding semi-discretizations are uniformly (with respect to $h > 0$) null controllable. Convergence of controls and states holds as well.

In this case the most natural functional setting is the following one. The initial data $y^0$ belongs to $L^2(\Omega)$, the control $f$ lies in $L^2(\omega \times (0, T))$ and the solutions then belong to $C([0, T]; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$. Convergences hold in these classes as well.

**3.3. General domains.** The methods of proof of the previous section, based on Fourier series expansions, do not apply to general domains. In fact, even in the context of the continuous heat equation, the existing proofs of null controllability require obtaining the observability estimates by Carleman inequalities (see [5] and [11]). So far the discrete or semi-discrete version of these Carleman inequalities

and its possible applications to observability estimates for numerical approximation schemes for the heat equation is a completely open subject of research.

However, in view of the results of the previous section, and using a classical argument, based on extending the control domain and then getting the controls as restrictions to the original boundary of the controlled states, one can derive similar results for general domains but provided *the controls are supported everywhere on the boundary of the domain.* The problem of determining sharp conditions on the subsets of the boundary so that the semi-discrete systems are uniformly controllable is completely open. As we have mentioned above, even in the simplest geometry of the square domain of the previous subsection, the result fails to hold without some restrictions on the support of the control that are not needed for the continuous heat equation.

The following holds:

**Theorem 3.5.** *For all bounded smooth domain $\Omega$, all time $T > 0$ and all initial data $y^0 \in L^2(\Omega)$, there exists a uniformly bounded sequence of discrete controls $V_h \in L^2(\partial\Omega_h \times (0,T))$ ensuring the null controllability of the finite-difference semi-discrete approximation in $\Omega_h$. These controls can be chosen so that the solutions $Y_h$ converge weakly in $L^2(0,T;H^1(\Omega))$ to the solution $y$ of the heat equation satisfying the null final condition* (3.2).

*Proof.* Let us briefly explain how this classical extension-restriction method can be implemented in this framework.

Without loss of generality we can assume that $\Omega$ is contained in the square domain $\tilde{\Omega} = (0,\pi) \times (0,\pi)$. We discretize the square as in the previous sections, and define the approximating domains $\Omega_h$ as those that, having their boundary constituted by mesh-points, better approximate the domain $\Omega$. For the sake of simplicity we assume that $\Omega_h$ contains $\Omega$. We also consider a band-like control subdomain $\omega$ in the square $\tilde{\Omega}$ so that the results of the previous sections apply and $\Omega_h \cap \omega = \emptyset$ for all $h > 0$.

Given initial data $y^0 \in L^2(\Omega)$ for the continuous heat equation we define approximating discrete data $Y_h^0$ in $\Omega_h$ in a standard way, for instance, by simply taking on each mesh-point the average of $y^0$ on the neighboring square of sides of size $h$. This data can be easily extended by zero to discrete data $\tilde{Y}_h^0$ defined in the whole mesh of the square. In view of the results of the previous section (Remark 3.4 ) this generates controls $F_h$ with support in $\omega$, which are uniformly bounded in $L^2(\omega \times (0,T))$ and converging, as $h \to 0$, to the control of the heat equation (3.29) in the square $\tilde{\Omega}$. This yields also uniformly bounded states $\tilde{Y}_h$ in the space $C([0,T];L^2(\tilde{\Omega})) \cap L^2(0,T;H_0^1(\tilde{\Omega}))$. Obviously, here, as in previous sections, these bounds hold in fact for the piecewise linear continuous extensions of the discrete solutions.

More precisely, the corresponding solutions $\tilde{Y}_h$ converge to the solution $y$ of the heat equation in the space $L^2(0,T;H_0^1(\tilde{\Omega}))$. We can then restrict these solutions to the domains $\Omega_h$ and obtain the solutions $Y_h$ of the semi-discrete system in $\Omega_h$, which, by construction, satisfy the final null condition (3.13) and converge to the solution of the heat equation. These solutions satisfy non-homogeneous boundary conditions.

We read their trace as the boundary controls $V_h$ in $\partial\Omega_h$ (resp. $\partial\Omega$) for the semi-discrete (resp. continuous) heat equations. These controls are bounded in $L^2(\partial\Omega_h \times (0,T))$ because they are traces of solutions of bounded energy in $L^2(0,T;H_0^1(\tilde{\Omega}))$. Their weak convergence can also be proved. However, at this point one has to be careful since the controls are defined on boundaries $\partial\Omega_h$ that depend on $h$. A possible way of stating that convergence rigorously is considering smooth test functions $\theta(x)$ defined everywhere in the square and ensuring that $\int_{\partial\Omega_h} V_h\theta d\sigma$ tends to $\int_{\partial\Omega} v\theta d\sigma$, as $h \to 0$ for all smooth test functions $\theta$. This convergence property of controls holds as well. □

**Remark 3.6.** The method of proof we have presented based on the extension of the domains and using the previously proved results on the square has two main drawbacks:

- The first one is that the control is required to be supported everywhere on the boundary of the domain. We emphasize however that, despite the fact that no geometric restrictions are needed for the continuous heat equation, in the sense that null controllability holds from an arbitrarily small open subset of the boundary, that is not the case for the semi-discrete one. Thus, the class of subsets of the boundary for which passing to limit on the null-controllability property is possible is still to be clarified, and the result above showing that the whole boundary always suffices is the first positive one in this direction.

- The second one is that it is based on the results obtained in the square by Fourier series techniques. As we have mentioned above, the main tool to deal with continuous heat equations are the Carleman inequalities. As far as we know there is no discrete counterpart of those inequalities and this would be essential to deal with more general heat equations with variable coefficients, or semilinear perturbations. The methods described in Section 2 showing the two-grid controllability of the semilinear wave equation by compactness-uniqueness arguments do not apply for heat-like equations because of their very strong time-irreversibility. Thus, the Carleman approach seems to be the most promising one. However, the fact that observability fails for the semi-discrete system for some observation subdomains indicates that the problem is complex in the sense that the discrete version of the continuous Carleman inequality does not hold. This is a widely open subject of research.

**Remark 3.7.** Similar results hold for a semi-discrete regular finite-element approximation of the heat equation, as long as solutions can be developed in Fourier series, allowing to reduce the problem in the square to a one-parameter family of $1-d$ problems, and then apply the extension-restriction method to address general domains.

**Acknowledgements.** I am grateful to E. Trélat and X. Zhang for fruitful comments that contributed to improve the first version of this article. My thanks also go to

C. Simó for his valuable comments on the use of Newton's method for efficiently computing the controls obtained in Section 2 by means of fixed point techniques.

## References

[1] Cannarsa, P., Komornik, V., Loreti, P., Controllability of semilinear wave equations with infinitely iterated logarithms. *Control Cybernet.* **28** (3) (1999), 449–461.

[2] Castro, C., Micu, S., Boundary controllability of a linear semi-discrete 1-D wave equation derived from a mixed finite elements method. *Numer. Math.* **102** (3) (2006), 413–462.

[3] Coron, J.-M., Trélat, E., Global steady-state stabilization and controllability of 1-D semilinear wave equations. *Commun. Contemp. Math.*, to appear.

[4] Fernández-Cara, E., Zuazua, E., Null and approximate controllability for weakly blowing-up semilinear heat equations. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **17** (5) (2000), 583–616.

[5] Fursikov, A., Imanuvilov, O., *Controllability of evolution equations*. Lectures Notes Series 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, Seoul 1996.

[6] Glowinski, R., Ensuring well-posedness by analogy: Stokes problem and boundary control of the wave equation. *J. Comput. Phys.* **103** (1992), 189–221.

[7] Ignat, L., Zuazua, E., A two-grid approximation scheme for nonlinear Schrödinger equations: Dispersive properties and convergence. *C. R. Acad. Sci. Paris Sér. I Math.* **341** (6) (2005), 381–386.

[8] Infante, J. A., Zuazua, E., Boundary observability for the space-discretizations of the one-dimensional wave equation. *M2AN Math. Model. Numer. Anal.* **33** (2) (1999), 407–438.

[9] Komornik, V., Zuazua, E., A direct method for the boundary stabilization of the wave equation. *J. Math. Pures Appl.* **69** (1) (1990), 33–55.

[10] Labbé, S., Trélat, E., Uniform controllability of semidiscrete approximations of parabolic control systems. *Systems Control Lett.* **55** (7) (2006), 597–609.

[11] Lebeau, G., Robbiano, L., Contrôle exact de l'équation de la chaleur. *Comm. Partial Differential Equations* **20** (1995), 335–356.

[12] León, L., Zuazua, E., Boundary controllability of the finite-difference space semi-discretizations of the beam equation. *ESAIM Control Optim. Calc. Var.* **8** (2002), 827–862.

[13] Lions, J. -L., *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*. Tome 1, Contrôlabilité exacte, Rech. Math. Appl. 8, Masson, Paris 1988.

[14] López, A., Zuazua, E., Some new results related with the null-controllability of the $1-d$ heat equation. In *Séminaire sur les Equations aux Dérivées Partielles*, École Polytechnique, Palaiseau 1997–1998.

[15] López, A., Zuazua, E., Uniform null controllability for heat equations with rapidly oscillating coefficients. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **19** (5) (2002), 543–580.

[16] Loreti, P., Mehrenberger, M., A Ingham type proof for a bigrid observability theorem. Preprint #2005-012, Institut de Recherche Mathématique Avancée, Université Strasbourg, Strasbourg 2005; *ESAIM Control Optim. Calc. Var.*, to appear.

[17] Negreanu, M., Zuazua, E., Convergence of a multigrid method for the controllability of a 1-d wave equation. *C. R. Acad. Sci. Paris Sér. I Math.* **338** (4) (2004), 413–418.

[18] Ramdani, K., Takahashi, T., Tucsnak, M., Uniformly exponentially stable approximations for a class of second order evolution equations. Preprint 27/2003, Institut Elie Cartan, Nancy 2003; *ESAIM Control Optim. Calc. Var.*, to appear.

[19] Tcheugoue, L. R., Zuazua, E., Uniform exponential long time decay for the space semi-discretizations of a damped wave equation with artificial numerical viscosity. *Numer. Math.* **95** (3) (2003), 563–598.

[20] Tcheugoue, L. R., Zuazua, E., Uniform boundary stabilization of the finite difference space discretization of the $1-d$ wave equation. *Adv. Comput. Math.*, to appear.

[21] Trefethen, L. N., Group velocity in finite difference schemes. *SIAM Rev.* **24** (2) (1982), 113–136.

[22] Zhang, X., Explicit observability estimate for the wave equation with potential and its application. *Royal Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **456** (2000), 1101–1115.

[23] Zhang, X., Zuazua, E., Exact controllability of the semi-linear wave equation. In *Sixty Open Problems in the Mathematics of Systems and Control* (ed. by V. D. Blondel and A. Megretski), Princeton University Press, Princeton, N.J., 2004, 173–178.

[24] Zuazua, E., Exact controllability for the semilinear wave equation. *J. Math. Pures Appl.* **69** (1) (1990), 1–32.

[25] Zuazua, E., Exact boundary controllability for the semilinear wave equation. In *Nonlinear Partial Differential Equations and their Applications* (ed. by H., Brezis and J.-L. Lions), Collège de France Seminar, Vol. X, Pitman Res. Notes Math. Ser. 220, Longman Scientific & Technical, Harlow 1991, 357–391.

[26] Zuazua, E., Exact controllability for the semilinear wave equation in one space dimension. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **10** (1993), 109–129.

[27] Zuazua, E., Null controllability of the heat equation in thin domains. In *Equations aux dérivées partielles et applications* (Articles dédiés à Jacques-Louis Lions), Gauthier-Villars, Paris 1998, 787–801.

[28] Zuazua, E., Boundary observability for the finite-difference space semi-discretizations of the $2-d$ wave equation in the square. *J. Math. Pures Appl.* **78** (1999), 523–563.

[29] Zuazua, E., Optimal and approximate control of finite-diffference schemes for the 1D wave equation. *Rend. Mat. Appl.* (7) **24** (2) (2004), 201–237.

[30] Zuazua, E., Propagation, observation, and control of waves approximated by finite difference methods. *SIAM Rev.* **47** (2) (2005), 197–243.

[31] Zuazua, E., Controllability and Observability of Partial Differential Equations: Some results and open problems. In *Handbook of Differential Equations: Evolutionary Differential Equations* (ed. by Dafermos, C., Feireisl, E.), vol. 3, Elsevier Science, to appear.

Departamento de Matemáticas, Universidad Autónoma, 28049 Madrid, Spain
E-mail: enrique.zuazua@uam.es

# Multiscale modeling for epitaxial growth

Russel E. Caflisch*

**Abstract.** Epitaxy is the growth of a thin film on a substrate in which the crystal properties of the film are inherited from those of the substrate. Because of the wide range of relevant length and time scales, multiscale mathematical models have been developed to describe epitaxial growth. This presentation describes atomistic, island dynamics and continuum models. Island dynamics models are multiscale models that use continuum coarse-graining in the lateral direction, but retain atomistic discreteness in the growth direction. Establishing connections between the various length and time scales in these models is a principal goal of mathematical materials science. Progress towards this goal is described here, including the derivation of surface diffusion, line tension and continuum equations from atomistic, kinetic models.

**Mathematics Subject Classification (2000).** Primary 82D25; Secondary 82C24.

**Keywords.** Epitaxial growth, island dynamics, step edge, step stiffness, Gibbs–Thomson, adatom diffusion, line tension, surface diffusion, renormalization group, kinetic Monte Carlo.

## 1. Introduction

Epitaxy is the growth of a thin film on a substrate in which the crystal properties of the film are inherited from those of the substrate. Since an epitaxial film can (at least in principle) grow as a single crystal without grain boundaries or other defects, this method produces crystals of the highest quality. In spite of its ideal properties, epitaxial growth is still challenging to mathematically model and numerically simulate because of the wide range of length and time scales that it encompasses, from the atomistic scale of Ångstroms and picoseconds to the continuum scale of microns and seconds.

The geometry of an epitaxial surface consists of step edges and island boundaries, across which the height of the surface increases by one crystal layer, and adatoms which are weakly bound to the surface. Epitaxial growth involves deposition, diffusion and attachment of adatoms on the surface. Deposition is from an external source, such as a molecular beam. Figure 1 provides a schematic illustration of the processes involved in epitaxial growth.

The models that are most often used to describe epitaxial growth include the following: A typical *Kinetic Monte Carlo* (KMC) method simulates the dynamics of

*This work was partially supported by the National Science Foundation through grant DMS-0402276 and by the Army Research Office through grant DAAD19-02-1-0336

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

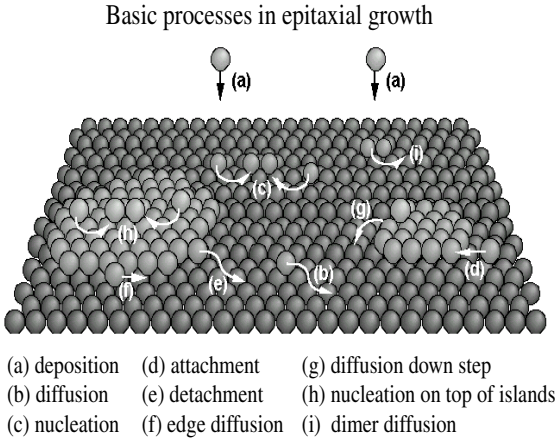



Figure 1. Schematic view of the processes involved in epitaxial growth.

the epitaxial surface through the hopping of adatoms along the surface. The hopping rate has the Arrhenius form $e^{-E/kT}$ in which $E$ is the energy barrier for going from the initial to the final position of the hopping atom. *Island dynamics* describe the surface through continuum scaling in the lateral directions but atomistic discreteness in the growth direction. *Continuum equations* approximate the surface using a smooth height function $z = h(x, y, t)$, obtained by coarse graining in all directions. Two other models are used to describe epitaxial growth on a limited time range. *Molecular dynamics* (MD) consists of Newton's equations for the motion of atoms on an energy landscape. Because the time scale for MD is femtoseconds ($10^{-15}$ seconds), this can only be applied to very short time periods. *Rate equations* describe the surface through a set of bulk variables without spatial dependence. With some exceptions [5], [16], these have been used only for submonolayer growth.

The initial theories for epitaxial growth, such as [3], relied on an assumption that the system is close to equilibrium. In many epitaxial systems, however, the system is far from equilibrium so that a kinetic description is required. The emphasis in this article will be on KMC, island dynamics and continuum models for epitaxial systems that are far from equilibrium. A principal goal of mathematical materials science is to analyze the connections between these models. The results presented below, from the work of Margetis [18], Caflisch & Li [7], Margetis & Caflisch [19], Chua et al. [10], and Haselwandter & Vvedensky [14], [15] are for surface diffusion and step stiffness derived from atomistic kinetic models of epitaxy, and general continuum equations from a simplified model. These results are among the first of their kind; e.g., the formula for step stiffness comes from the first derivation of the Gibbs–Thomson formula from an atomistic, kinetic model rather than from a thermodynamic driving force. The results are far from complete. Other effects, such as the nonlinear terms in the continuum equations, have not been derived for a full model of epitaxy. In

addition, the derivations presented here are based on formal asymptotics, rather than rigorous mathematical analysis. Nevertheless, these results are a significant step toward a more complete theory and can serve as a starting point for more rigorous analysis.

For simplicity in the presentation, the lattice constant $a$ will be taken to be $a = 1$, except in a few places where it is useful as a placeholder. Also, all transition rates (with units 1/time) are expressed in terms of equivalent diffusion constants (with units length$^2$/time); i.e., a rate $r$ is replaced by a diffusion coefficient $D = a^2 r$.

## 2. Mathematical models for epitaxial growth

In this section, various models for epitaxial growth are described, including atomistic KMC, island dynamics and continuum models, as well as a kinetic model for the structure of a step edge (or island boundary) that is used with island dynamics.

**2.1. Atomistic Models.** The simplest KMC model is a simple cubic pair-bond solid-on-solid (SOS) model [28], [29]. In this model, there is a stack of atoms, without vacancies, above each site on a two-dimensional lattice. New atoms are randomly deposited at a deposition rate $F$. Any surface atom (i.e., the top atom in the stack of atoms at a lattice point) is allowed to move to its nearest neighbor site at a rate $r = D/a^2$ in which $a$ is the lattice constant and $D$ is a diffusion coefficient. In the simplest case, $D$ is determined by

$$D = D_0 \exp\{-(E_S + nE_N)/k_B T\}. \tag{1}$$

In this equation, $D_0$ is a constant prefactor of size $10^{13} a^2 s^{-1}$, $k_B$ is the Boltzmann constant, $T$ is the surface temperature, $E_S$ and $E_N$ represent the surface and nearest neighbor bond energies, and $n$ is the number of in-plane nearest neighbors. The terrace diffusion coefficient $D_T$ for adatoms on a flat terrace and the edge diffusion coefficient $D_E$ for adatoms along a step edge (with a single in-plane neighbor) are

$$D_T = D_0 \exp\{-E_S/k_B T\}, \tag{2}$$

$$D_E = D_0 \exp\{-(E_S + E_N)/k_B T\}. \tag{3}$$

Validity of this KMC model for epitaxial growth has been demonstrated by comparison to RHEED measurements from molecular beam epitaxy (MBE) experiments [11]. More complicated models for the diffusion coefficient, subject to the condition of detailed balance, are also used.

**2.2. Island dynamics models.** Burton, Cabrera and Frank [3] developed the first detailed theoretical description for epitaxial growth. This BCF model is an "island dynamics" model, since it describes an epitaxial surface by the location and evolution of the island boundaries and step edges. It employs a mixture of coarse graining and

atomistic discreteness, since island boundaries are represented as smooth curves that signify an atomistic change in crystal height.

Adatom diffusion on the epitaxial surface is described by a diffusion equation of the form

$$\partial_t \rho - D_T \nabla^2 \rho = F - 2(dN_{\text{nuc}}/dt) \tag{4}$$

in which $F$ is the deposition flux rate and the last term represents loss of adatoms due to nucleation. Desorption from the epitaxial surface has been neglected.

The net flux to the step edge from upper and lower terraces is denoted as $f_+ = f_+(y,t)$ and $f_- = f_-(y,t)$, respectively, in which

$$v\rho_+ + D_T \boldsymbol{n} \cdot \nabla \rho_+ = -f_+, \tag{5}$$

$$v\rho_- + D_T \boldsymbol{n} \cdot \nabla \rho_- = f_-. \tag{6}$$

The total flux is

$$f = f_+ + f_-. \tag{7}$$

Different island dynamics models are distinguished by having different formulas for the diffusive boundary conditions and normal velocity.

1. *The island dynamics model with irreversible aggregation:*

$$\begin{aligned} \rho &= 0, \\ v &= f. \end{aligned} \tag{8}$$

2. *The BCF boundary conditions:*

$$\begin{aligned} \rho &= \rho_*, \\ v &= f, \end{aligned} \tag{9}$$

   in which $\rho_*$ is the equilibrium adatom density at a step.

3. *The island dynamics model with step-edge kinetics:*

$$\begin{aligned} f_+ &= (D_T \rho_+ - D_E \phi)\cos\theta, \\ f_- &= (D_T \rho_- - D_E \phi)\cos\theta, \\ v &= k\, w \cos\theta, \end{aligned} \tag{10}$$

   in which $\phi$ and $k$ are the densities of edge-atoms and kinks, and $w$ is the kink velocity, defined in Section 2.3.

4. *The island dynamics model with line tension and surface diffusion:*

$$\begin{aligned} f_+ &= D_{d+}(\rho_+ - \rho_*) - \mu_+ \kappa, \\ f_- &= D_{d-}(\rho_- - \rho_*) - \mu_- \kappa, \\ v &= D_T \boldsymbol{n} \cdot [\nabla \rho] + \beta \rho_{*yy} + (\mu/D_E)\kappa_{ss}, \end{aligned} \tag{11}$$

in which $\kappa$ is curvature and $\kappa_{ss}$ is its second derivative along the length of a step edge, $\rho_*$ is a reference adatom density, $D_{d\pm}$ are the attachment/detachment rates, and $\mu_\pm = (D_{d\pm}\rho_*/k_BT)\tilde{\gamma}$ in which $\tilde{\gamma}$ is the step stiffness. This is further discussed in Section 4.

For the case of irreversible aggregation, a dimer (consisting of two atoms) is the smallest stable island, and the nucleation rate is

$$\frac{dN_{\text{nuc}}}{dt} = D\sigma_1\langle\rho^2\rangle, \tag{12}$$

where $\langle\cdot\rangle$ denotes the spatial average of $\rho(\boldsymbol{x}, t)^2$ and

$$\sigma_1 = \frac{4\pi}{\ln[(1/\alpha)\langle\rho\rangle D/F]} \tag{13}$$

is the adatom capture number. The parameter $\alpha$ reflects the island shape, and $\alpha \simeq 1$ for compact islands. Expression (12) for the nucleation rate implies that the time of a nucleation event is chosen deterministically. The choice of the location of the new island is determined by probabilistic choice with spatial density proportional to the nucleation rate $\rho^2$. This probabilistic choice constitutes an atomistic fluctuation that is retained in the island dynamics model [24].

Snapshots of the results from a typical island dynamics simulation are shown in Figure 2. Shown is the island geometry after coverage of 0.25 layers (left) and

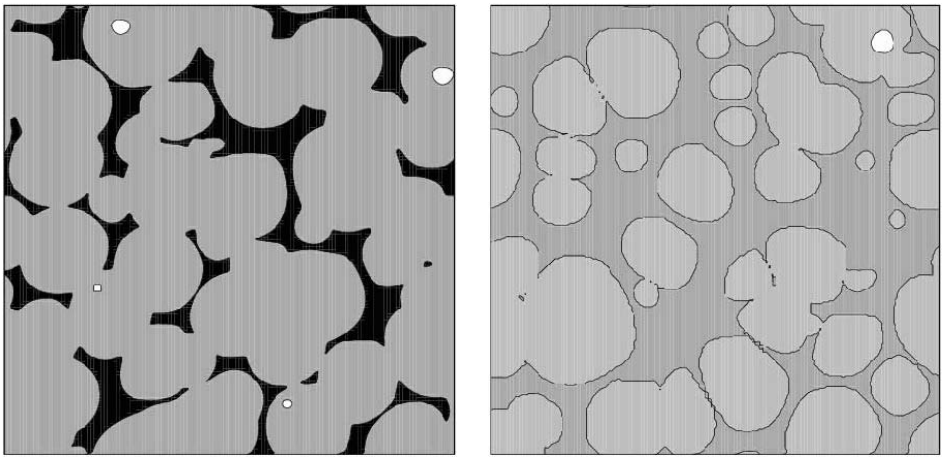


Figure 2. Island geometry for island dynamics with irreversible aggregation after deposition of 0.25 layers (left) and 10.25 layers (right).

coverage of 10.25 layers (right). These simulations are for irreversible aggregation with boundary conditions from Eq. (8). Numerical simulation of the island dynamics is performed using a level set method for thin film growth, as described in [4], [8].

Validation of the island dynamics/level set method has been performed by careful comparison to the results of the atomistic KMC models. Various generalizations and additional physics results are described in [22], [23]. Related work on level set methods for epitaxial growth are found in [9], [25], [26].

The principal dimensionless parameters for epitaxial growth are the ratios of flux and diffusive coefficients, which we refer to as "Péclet numbers" by analogy with fluid mechanics. Let $\bar{f}$ be a characteristic size for the flux to an edge. Let $P_T$ be the terrace Péclet number and $P_E$ be the edge Péclet number, defined as

$$P_T = F/D_T, \tag{14}$$

$$P_E = f/D_E, \tag{15}$$

in which $D_E$ is the edge diffusion constant. Typical values for $P_T^{-1} = D_T/F$ are in the range of $10^4$ to $10^8$.

**2.3. The kinetic edge model.** The kinetic edge model of island dynamics was developed in [6]. It involves a statistical description of the crystalline structure of a step edge, including the edge-atom density $\phi$ and the kink density $k$. Edge-atoms are atoms with a single in-plane neighbor along the step; kinks are atoms with two in-plane neighbors. Kinks are of two types – right-facing kinks and left-facing kinks – the densities of which are denoted by $k_r$ and $k_\ell$. Figure 3 provides a schematic picture

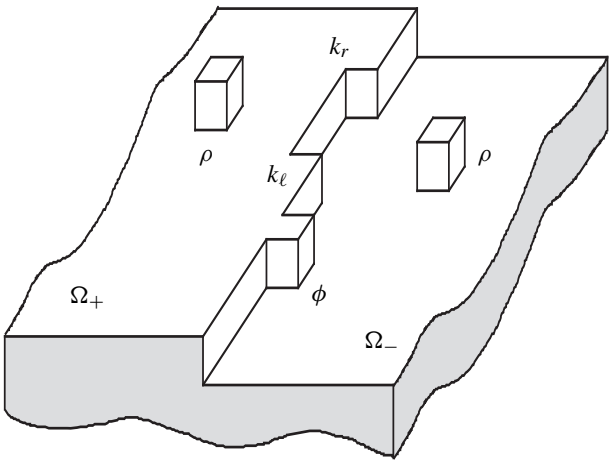



Figure 3. The geometry of step edge, showing adatoms with density $\rho$ on the upper and lower terraces $\Omega_+$ and $\Omega_-$, edge-atoms with density $\phi$, and right and left facing kinks with density $k_r$ and $k_\ell$.

of the kink density model for a step edge. Related models have been derived by Balykov et al. [1], Balykov & Voigt [2] and Filimonov & Hervieu [13].

The kinetic edge model consists of a diffusion equation for the edge-atom density $\phi$ and a convection equation for the kink density $k$

$$\partial_t \phi - D_E \partial_s^2 \phi = f_+ + f_- - f_0, \tag{16}$$

$$\partial_t k + \partial_s (w(k_r - k_\ell)) = 2(g - h). \tag{17}$$

In Eq. (16), $f_\pm$ are the net fluxes to the edge from terraces as defined in Eq. (5) and Eq. (6), and $f_0$ is the net loss term due to the attachment of edge-atoms to kinks. In Eq. (17), $w$ is the kink velocity, and $g$ and $h$ represent, respectively, the creation and annihilation of left-right kink pairs. Note that left-facing kinks and right-facing kinks move in opposite directions with velocity $w$ and $-w$, respectively. The total kink density and the relation between the kink density and the normal angle [3] are

$$k_r + k_\ell = k, \tag{18}$$

$$k_r - k_\ell = \tan\theta. \tag{19}$$

The quantities $f_+$, $f_-$, $f_0$, $w$, $g$, $h$, and $v$ are determined by the following constitutive relations (in simplified form):

$$f_+ = (D_T \rho_+ - D_E \phi)\cos\theta, \tag{20}$$

$$f_- = (D_T \rho_- - D_E \phi)\cos\theta, \tag{21}$$

$$f_0 = v\,(\phi\kappa + 1)\,, \tag{22}$$

$$w = l_1 D_E \phi + D_T(l_2\rho_+ + l_3\rho_-) = l_{123} D_E \phi + (l_2 f_+ + l_3 f_-)/\cos\theta, \tag{23}$$

$$\begin{aligned} g &= \phi\,(m_1 D_E \phi + D_T(m_2\rho_+ + m_3\rho_-)) \\ &= \phi(m_{123} D_E \phi + (m_2 f_+ + m_3 f_-)/\cos\theta), \end{aligned} \tag{24}$$

$$\begin{aligned} h &= k_r k_\ell\,(n_1 D_E \phi + D_T(n_2\rho_+ + n_3\rho_-)) \\ &= k_r k_\ell (n_{123} D_E \phi + (n_2 f_+ + n_3 f_-)/\cos\theta), \end{aligned} \tag{25}$$

$$X_t = v = wk\cos\theta, \tag{26}$$

where $D_T$ is the (diffusion) hopping rate of an adatom on a terrace, $D_E$ is the (diffusion) hopping rate of an edge-atom along or off an edge, and all $l_i, m_i, n_i$ $(i = 1, 2, 3)$ are nonnegative numbers. The geometric parameters $l_i, m_i, n_i$ count the number of paths from one state to another, cf. [6] for details. Here, these parameters are generalized to allow a factor relating the macroscopic density $\rho$ or $\phi$ to the local density of adatoms or edge atoms at a specific site. For convenience, we have used the notation

$$q_{ij} = q_i + q_j \quad \text{and} \quad q_{ijk} = q_i + q_j + q_k$$

for $q = l, m$, or $n$. For simplicity in this presentation, the constitutive laws (20)–(25) have been simplified by omission of terms that are insignificant for the kinetic steady state solutions of relevance to step-flow growth and by specialization to the case of $\theta$ near 0. The terms omitted from (20)–(25) include terms that are important for detailed

balance, so that they are required for determination of the equilibrium solution for this model. In the more complete analysis of [6], [7], [19], all of the neglected terms are included.

There are several significant solutions for the kinetic step edge model Eq. (16)–Eq. (26). First, there is an equilibrium solution that was originally determined in [3] (note that some terms that have been omitted from the presentation here are significant for the equilibrium). Second there is a kinetic (i.e., nonequilibrium) steady state solution, for which the presentation includes all of the significant terms. Suppose that the kink density and edge Peclet number are small (i.e., $ak \ll 1$ and $P_E \ll 1$) and that the step is symmetric (i.e., $\rho_+ = \rho_-$), then the adatom, edge-adatom and kink densities of the kinetic steady state are approximately

$$\rho = (D_E/D_T)a^{-1}\varphi, \tag{27}$$

$$\varphi = (16a/3)k^2, \tag{28}$$

$$k = \left(\frac{16}{15}P_E\right)^{\frac{1}{3}} a^{-1}. \tag{29}$$

The exponent $1/3$ in (29) is related to the critical size for formation of a left-right kink pair. If the critical size were $j$ (i.e., if $j+1$ edge-adatoms were required to form a stable kink pair) then the exponent would be $j/(j+2)$.

Figure 4 shows a comparison of this steady state solution (solid line) and computational results from KMC (squares, circles and triangles) for kink density $k$. The

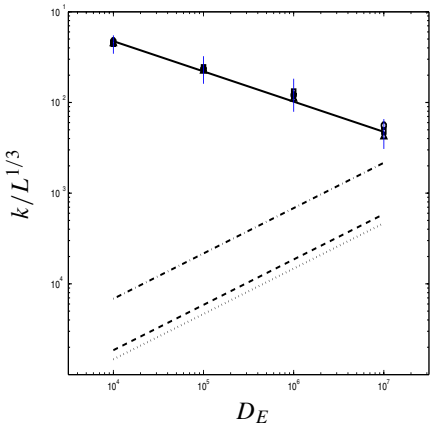



Figure 4. Kink density $k$, normalized by $L^{1/3}$, vs. edge diffusion coefficient $D_E$ for kinetic steady state, for various values of terrace width $L$. Parameter values are flux $F = 1$ and adatom diffusion $D_T = 10^{12}$. Results are shown from the kinetic theory (solid line) and KMC computations with $L = 25$ (squares), $L = 50$ ($\circ$), and $L = 100$ ($\triangle$). These are compared to the corresponding equilibrium values for $L = 25$ (dash-dotted line), $L = 50$ (dashed line), and $L = 100$ (dotted line), showing that the steady state and equilibrium differ both qualitatively and quantitatively.

BCF equilibrium values for $k$ are also plotted (lower three lines) for comparison. In this figure, $F = 1$ and $D_T = 10^{12}$, while $D_E$ varies between $10^4$ and $10^7$. The computations are for a periodic step train with straight steps at angle $\theta = 0$ and with distance $L = 25$, 50, and 100 between the steps. The figure shows excellent agreement between the predictions of the present theory and the results of the KMC simulation, with differences that are less than one standard deviation of the KMC results. The results are significantly different from equilibrium both in size and in dependence on $D_E$.

**2.4. Continuum models.** Continuum models of epitaxial growth employ coarse graining in all directions. In most cases, they describe the epitaxial surface through a smooth height function $z = h(x, y, t)$. The general equation of this type, as discussed by Haselwandter and Vvedensky [14], [15], is

$$h_t = \nu_2 \nabla^2 h - \nu_4 \nabla^4 h + \lambda_{13} \nabla (\nabla h)^3 + \lambda_{22} \nabla^2 (\nabla h)^2 + \xi \tag{30}$$

in which the $\nu_2$ term comes from a height-dependence in the energy, the $\nu_4$ term is surface diffusion, the $\lambda$ terms are nonlinearities, and $\xi$ is a stochastic noise term. This equation generalizes previous models, including the Edwards–Wilkinson (EW) equation with $\nu_4 = \lambda_{13} = \lambda_{22} = 0$, the Mullins–Herring (MH) equation with $\nu_2 = \lambda_{13} = \lambda_{22} = 0$, and the Villain–Lai–Das Sarma (VLDS) equation with $\nu_2 = \lambda_{13} = 0$. The relations between these models are further discussed in Section 5, through a renormalization group analysis applied to the Edwards–Wilkinson and Wolf–Villain models for epitaxial growth.

Derivation of these continuum equations has been mostly through symmetry arguments, thermodynamic driving force or by heuristics. The results reported in Sections 3 and 5 are among the first derivation of these equations from kinetic, atomistic models. Alternative modeling approaches have included additional dependent variables, not just the interface height. For example, Lo & Kohn [17] included adatom density in addition to height. Margetis et al. [20] derive similar results starting from an island dynamics model with a kinetic step edge description, as in Section 2.2 and 2.3.

## 3. Surface diffusion

As first derived by Mullins [21], the surface diffusion equation

$$h_t = -\nu_4 \nabla^4 h \tag{31}$$

describes the evolution of a surface through diffusion of the material that comprises the surface. Margetis [18] has given an atomistic, kinetic derivation of surface diffusion for epitaxial growth, and he found the surprising result that surface diffusion is not isotropic. While his derivation is based on detailed asymptotics starting from an island

dynamics model, the presentation here will be phenomenological but faithful to the spirit of Margetis's derivation.

Consider an epitaxial surface that consists of a series of steps that are nearly parallel. The terrace width $\ell$ between steps is approximately $\ell = a/|\nabla h|$ in which $a$ is the lattice constant. Assume that the steps are slowly varying but widely spaced, so that $a \ll \ell \ll \lambda$ in which $\lambda$ is the length scale for the variation in the steps away from straight. Also, assume that edge diffusion coefficient $D_E$ and the edge attachment/detachment rate $D_d$ are much smaller than the terrace diffusion coefficient $D_T$.

The analysis is based on the following fundamental property of diffusion: Consider a composite consisting of strips of two material with diffusion coefficients $D_1$ and $D_2$ and with strip widths $a_1$ and $a_2$. The effective diffusion coefficient $D_*$ for the composite is the arithmetic average if the diffusion is in the direction along the strips (i.e., a parallel configuration) and it is the harmonic average if the diffusion is in the direction perpendicular to the strips (i.e., a series configuration); i.e.,

$$D_* = \begin{cases} (a_1 D_1 + a_2 D_2)/(a_1 + a_2) & \text{parallel configuration,} \\ ((a_1 D_1^{-1} + a_2 D_2^{-1})/(a_1 + a_2))^{-1} & \text{series configuration.} \end{cases} \tag{32}$$

Define a tangential variable $s$ along the steps and a normal variable $n$ perpendicular to the steps. In the tangential direction, adatoms diffuse at the terrace diffusion rate of $D_T$ on the terraces between steps and at the edge diffusion rate $D_E$ along the steps. Since the terraces and steps are in parallel in the tangential direction, the corresponding diffusion coefficient

$$D_{ss} = (aD_E + \ell D_T)/(a + \ell) \approx D_T. \tag{33}$$

Diffusion of adatoms normal the steps is also at rate $D_T$, but it is interrupted by attachment and detachment from the steps at rate $D_d$. Since the terraces and steps are in a series configuration in the normal direction $n$, the diffusion coefficient in this direction is

$$D_{nn} = ((2aD_d^{-1} + \ell D_T^{-1})/(a + \ell))^{-}1 \approx D_T(1 + m|\nabla h|)^{-1} \tag{34}$$

in which

$$m = 2D_T/D_d. \tag{35}$$

The factor of 2 in the last two formulas is due to the details of the attachment/detachment model used in [18].

Now follow the derivation of diffusion from the thermodynamic driving force (but note that Margetis used a perturbation expansion based on the kinetic equations rather than this near-equilibrium argument). The evolution of the height $h$ is given in terms of the mobility tensor $M$, current $j$ and chemical potential $\mu$ as

$$\begin{aligned} h_t &= -\nabla \cdot j \\ &= \nabla \cdot (M \nabla \mu) \end{aligned} \tag{36}$$

since $j = -M\nabla\mu$ and $\mu = \delta E/\delta h = -g_1\nabla\cdot(\nabla h/|\nabla h|) - g_3\nabla h\cdot(|\nabla h|\nabla h)$. By the argument above, $M$ is the matrix

$$M = \frac{D_T\rho_*}{K_BT}\begin{pmatrix}1 & 0\\ 0 & (1+m|\nabla h|)^{-1}\end{pmatrix} \tag{37}$$

in the $n$, $s$ coordinates.

## 4. Step stiffness

In [7], Caflisch and Li considered the zero limit of the edge Peclet $P_E$ number for the kinetic edge model from Section 2.3 for a step that is a slight perturbation of a straight step with $\theta = 0$, i.e., parallel to a crystallographic direction. They used a very specific form for the wavelength and amplitude of the perturbation and their scaling with $P_E$, and they assumed that the solution was close to the kinetic steady state Eq. (29) for $\theta = 0$. They derived the boundary conditions Eq. (11) for the evolution of a step, including the Gibbs–Thomson form of the step stiffness and a term due to edge diffusion of the adatoms that attach to the step.

More recently, Margetis and Caflisch [19] performed a more general analysis for a step with variable $\theta$. Under the assumption that $P_E \ll 1$ and that the solution is close to the kinetic steady state, they identified several regimes for the behavior of the solution and the step stiffness coefficient. Since the complete results are complicated, we present the results in an abbreviated form and refer to [19] for a detailed expression.

First, suppose that the curvature $\kappa$ of the step satisfies

$$|\kappa| < O(P_E) \ll 1. \tag{38}$$

Then the step edge kinetics allow for two regimes for the step stiffness $\tilde\gamma$:

$$\tilde\gamma = \begin{cases}(k_BT/D_T\rho_*)\theta^{-1} & \text{for } P_E^{1/3} \ll \theta \ll 1,\\ (k_BT/D_T\rho_*)\tilde\gamma_0 & \text{for } 0 < \theta \ll P_E^{1/3}.\end{cases} \tag{39}$$

The results for step stiffness imply results for the line tension $\gamma$, since $\tilde\gamma = \gamma + \gamma_{\theta\theta}$ then to leading order

$$\gamma \approx \begin{cases}(kT_B/D_T\rho_*)\theta\log\theta & \text{for } P_E^{1/3} \ll \theta \ll 1,\\ c_0 & \text{for } 0 < \theta \ll P_E^{1/3}.\end{cases} \tag{40}$$

in which $c_0$ is an undetermined constant. In the outer solution, $\gamma_\theta$ is nearly infinite for $\theta$ small, which predicts a flat facet corresponding to $\theta = 0$. The inner solution provides some curvature to this facet, however. These results are consistent with the recent results of Stasevich et al. [27] for the step stiffness in an Ising model.

## 5. Coarse graining

In a remarkable series of papers [10], [14], [15], Chua et al. and Haselwandter & Vvedensky performed coarse-graining, followed by a renormalization group analysis, for the Edwards–Wilkinson [12] and Wolf–Villain [30] models of epitaxial growth. In the Wolf–Villain model, particles are randomly deposited on the surface at rate $F$. Instead of diffusing along the surface, however, each deposited particle makes a single hop to the nearest neighbor site that has the largest number of neighbors (i.e., the highest coordination), or stays where it landed if that hop does not increase the number of neighbors. In two-dimensions (i.e., a one dimensional surface), a particle hops to a position that is no higher than its original position, while in three-dimensions, some particles may increase their coordination by hopping to a higher position.

For a general lattice model, Haselwandter and Vvedensky first write the stochastic evolution in terms of a Chapman–Kolmogorov transition probability $T_t(H_2|H_1)$ for transition between height configuration $H_1$ to $H_2$ in time $t$

$$T_{t+t'}(H_3|H_1) = \sum_{H_2} T_{t'}(H_3|H_2)T_t(H_2|H_1). \tag{41}$$

This can be converted to a Master equation

$$P(H,t) = \sum_r [W((H-r;r)P(H-r,t) - W(H;r)P(H,t)] \tag{42}$$

in which $P(H,t)$ is the probability for height configuration $H$ at time $t$, $W(H;r)$ is the transition rate between $H$ and $H+r$, and $r$ is the array of jump lengths between configurations. They then apply a Kramers–Moyal–van Kampen expansion with "largeness" parameter $\Omega$, with the lattice size and time between depositions being proportional to $\Omega^{-1}$. In the limit $\Omega \to \infty$, this expansion yields a lattice Langevin equation

$$\partial h_{ij}/\partial t = K_{ij}^{(1)} + \eta_{ij} \tag{43}$$

in which $K_{ij}^{(1)}$ are the first moment of the transition rates and $\eta_{ij}$ are Gaussian noises with mean zero and covariances given by the second moment of the transition rates. After performing a smoothing and a cutoff Taylor expansion and specializing to the Wolf–Villain (or Edwards–Wilkinson) model, the Langevin equation becomes Eq. (30).

Finally they perform a renormalization group (RG) analysis of equation Eq. (30). In the RG "flow", length and time are scaled at an exponential rate in the flow variable $\ell$. This analysis shows that the RG fixed points consist of the EW, MH and VLDS equations, as well as three previously unrecognized fixed points. The significance of this result is that as the solution of Eq. (30) evolves, it will linger near the fixed points, so that the solution will approximate a solution of each of these equations. The most important of the equations corresponding to these new fixed points, is their "FP1".

Although it is complicated in general, in the two dimensional case it has the form

$$h_t = -|\nu_2|\nabla^2 h - |\nu_4|\nabla^4 h - |\lambda_{13}|\nabla(\nabla h)^3 + \lambda_{22}\nabla^2(\nabla h)^2 + \xi \qquad (44)$$

which should be compared to Eq. (30). In two-dimensions this equation corresponds to a stable fixed point, but the corresponding equation in three dimensions corresponds to an unstable fixed point for the RG flow. This coarse graining and RG analysis provides both a derivation of these equations, starting from the EW or WV model, as well as an indication of the regimes of their validity.

**Acknowledgments.** The presentation here benefitted from discussions with Ted Einstein, Christoph Haselwandter, Dionisios Margetis, Tim Stasevich, Axel Voigt, and Dimitri Vvedensky. In particular, Margetis, Haselwandter and Vvedensky provided the author with advance copies of their unpublished work.

## References

[1] Balykov, L., Kitamura, M., Maksimov, I. L., Nishioka, K., Kinetics of non-equilibrium step structure. *Phil. Mag. Lett.* **78** (1998), 411–418.

[2] Balykov, L., Voigt, A., Kinetic model for step flow growth of [100] steps. *Phys. Rev. E* **72** (2005), #022601.

[3] Burton, W. K., Cabrera, N., Frank, F. C., The growth of crystals and the equilibrium structure of their surfaces. *Phil. Trans. Roy. Soc. London Ser. A* **243** (1951), 299–358.

[4] Caflisch, R. E., Gyure, M. F., Merriman, B., Osher, S., Ratsch, C., Vvedensky, D. D., Zinck, J. J., Island dynamics and the level set method for epitaxial growth. *Appl. Math. Lett.* **12**, 13 (1999), 13–22.

[5] Caflisch, R. E., Meyer, D., A reduced order model for epitaxial growth. *Contemp. Math.* **303** (2002), 9–23.

[6] Caflisch, R. E., E, W., Gyure, M. F., Merriman, B., Ratsch, C., Kinetic model for a step edge in epitaxial growth. *Phys. Rev. E* **59** (1999), 6879–6887.

[7] Caflisch, R. E., Li, B., Analysis of island dynamics in epitaxial growth. *Multiscale Model. Sim.* **1** (2002), 150–171.

[8] Chen, S., Kang, M., Merriman, B., Caflisch, R. E., Ratsch, C., Fedkiw, R., Gyure, M. F., Osher, S., Level set method for thin film epitaxial growth. *J. Comp. Phys.* **167** (2001), 475–500.

[9] Chopp, D., A level-set method for simulating island coarsening. *J. Comp. Phys.* **162** (2000), 104–122.

[10] Chua, A. L.-S., Haselwandter, C. A., Baggio, C., Vvedensky, D. D., Langevin equations for fluctuating surfaces. *Phys. Rev. E* **72** (2005), #051103.

[11] Clarke, S., Vvedensky, D. D., Origin of reflection high-energy electron-diffraction intensity oscillations during molecular-beam epitaxy: A computational modeling approach. *Phys. Rev. Lett.* **58** (1987), 2235–2238.

[12] Edwards, S. F., Wilkinson, D. R., The surface statistics of a granular aggregate. *Proc. Roy. Soc. Ser. A* **381** (1982), 17–31.

[13] Filimonov, S. N., Hervieu, Y. Y., Terrace-edge-kink model of atomic processes at the permeable steps. *Surface Sci.* **553** (2004), 133–144.

[14] Haselwandter, C. A., Vvedensky, D. D., Multiscale theory of fluctuating interfaces during growth. Preprint, 2005.

[15] Haselwandter, C. A., Vvedensky, D. D., Multiscale theory of fluctuating interfaces: From self-affine to unstable growth. Preprint, 2005.

[16] Kariotis, R., Lagally, M. G., Rate equation modelling of epitaxial growth. *Surface Sci.* **216** (1989), 557–578.

[17] Lo, T. S., Kohn, R. V., A new approach to the continuum modeling of epitaxial growth: slope selection, coarsening, and the role of the uphill current. *Physica D* **161** (2002), 237–257.

[18] Margetis, D., Unified continuum approach to crystal surface morphological relaxation. Preprint, 2005.

[19] Margetis, D., Caflisch, R. E., Modified kinetic model of step edge dynamics: modeling and derivation of step stiffness. Preprint, 2005.

[20] Margetis, D., Voigt, A., Caflisch, R. E., private communication, 2005.

[21] Mullins, W., Theory of thermal grooving. *J. Appl. Phys.* **28** (1957), 333–339.

[22] Petersen, M., Ratsch, C., Caflisch, R. E., Zangwill, A., Level set approach to reversible epitaxial growth. *Phys. Rev. E* **64** (2001), #061602.

[23] Ratsch, C., Gyure, M., Caflisch, R. E., Gibou, F., Petersen, M., Kang, M., Garcia, J., Vvedensky, D. D., Level-set method for island dynamics in epitaxial growth. *Phys. Rev. B* **65** (2002), #195403.

[24] Ratsch, C., Gyure, M. F., Chen, S., Kang, M., Vvedensky, D. D., Fluctuations and scaling in aggregation phenomena. *Phys. Rev. B* **61** (2000), 10598–10601.

[25] Russo, G., Smereka, P., A level-set method for the evolution of faceted crystals. *SIAM J. Sci. Comput.* **21** (2000), 2073–2095.

[26] Smereka, P., Spiral crystal growth. *Physica D* **138** (2000), 282–301.

[27] Stasevich, T. J., Gebremariam, H., Einstein, T. L., Giesen, M., Steimer, C., Ibach H., Low-temperature orientation dependence of step stiffness on {111} surfaces. *Phys. Rev. B* **71** (2005), #245414.

[28] Vvedensky, D. D., Atomistic modeling of epitaxial growth: comparisons between lattice models and experiment. *Comp. Materials Sci.* **6** (1996), 182–187.

[29] Weeks, J. D., Gilmer, G. H., Dynamics of crystal growth. *Adv. Chem. Phys.* **40** (1979), 157–228.

[30] Wolf, D. E., Villain, J., Growth with surface diffusion. *Europhys. Lett.* **13** (1990), 389–394.

Mathematics Department, UCLA, Los Angeles, CA 90095-1555, U.S.A.

E-mail: Caflisch@math.ucla.edu

# Compressive sampling

Emmanuel J. Candès*

**Abstract.** Conventional wisdom and common practice in acquisition and reconstruction of images from frequency data follow the basic principle of the Nyquist density sampling theory. This principle states that to reconstruct an image, the number of Fourier samples we need to acquire must match the desired resolution of the image, i.e. the number of pixels in the image. This paper surveys an emerging theory which goes by the name of "compressive sampling" or "compressed sensing," and which says that this conventional wisdom is inaccurate. Perhaps surprisingly, it is possible to reconstruct images or signals of scientific interest accurately and sometimes even exactly from a number of samples which is far smaller than the desired resolution of the image/signal, e.g. the number of pixels in the image.

It is believed that compressive sampling has far reaching implications. For example, it suggests the possibility of new data acquisition protocols that translate analog information into digital form with fewer sensors than what was considered necessary. This new sampling theory may come to underlie procedures for sampling and compressing data simultaneously.

In this short survey, we provide some of the key mathematical insights underlying this new theory, and explain some of the interactions between compressive sampling and other fields such as statistics, information theory, coding theory, and theoretical computer science.

**Mathematics Subject Classification (2000).** Primary 00A69, 41-02, 68P30; Secondary 62C65.

**Keywords.** Compressive sampling, sparsity, uniform uncertainty principle, underdertermined systems of linear equations, $\ell_1$-minimization, linear programming, signal recovery, error correction.

## 1. Introduction

One of the central tenets of signal processing is the Nyquist/Shannon sampling theory: the number of samples needed to reconstruct a signal without error is dictated by its bandwidth – the length of the shortest interval which contains the support of the spectrum of the signal under study. In the last two years or so, an alternative theory of "compressive sampling" has emerged which shows that super-resolved signals and images can be reconstructed from far fewer data/measurements than what is usually considered necessary. The purpose of this paper is to survey and provide some of the key mathematical insights underlying this new theory. An enchanting aspect of compressive sampling it that it has significant interactions and bearings on some fields in the applied sciences and engineering such as statistics, information theory, coding

*The author is partially supported by an NSF grant CCF–515362.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

theory, theoretical computer science, and others as well. We will try to explain these connections via a few selected examples.

From a general viewpoint, sparsity and, more generally, compressibility has played and continues to play a fundamental role in many fields of science. Sparsity leads to efficient estimations; for example, the quality of estimation by thresholding or shrinkage algorithms depends on the sparsity of the signal we wish to estimate. Sparsity leads to efficient compression; for example, the precision of a transform coder depends on the sparsity of the signal we wish to encode [24]. Sparsity leads to dimensionality reduction and efficient modeling. The novelty here is that *sparsity has bearings on the data acquisition process itself, and leads to efficient data acquisition protocols.*

In fact, compressive sampling suggests ways to economically translate analog data into already compressed digital form [20], [7]. The key word here is "economically." Everybody knows that because typical signals have some structure, they can be compressed efficiently without much perceptual loss. For instance, modern transform coders such as JPEG2000 exploit the fact that many signals have a sparse representation in a fixed basis, meaning that one can store or transmit only a small number of adaptively chosen transform coefficients rather than all the signal samples. The way this typically works is that one acquires the full signal, computes the complete set of transform coefficients, encode the largest coefficients and discard *all* the others. This process of massive data acquisition followed by compression is extremely wasteful (one can think about a digital camera which has millions of imaging sensors, the pixels, but eventually encodes the picture on a few hundred kilobytes). This raises a fundamental question: because most signals are compressible, why spend so much effort acquiring all the data when we know that most of it will be discarded? Wouldn't it be possible to acquire the data in already compressed form so that one does not need to throw away anything? "Compressive sampling" also known as "compressed sensing" [20] shows that this is indeed possible.

This paper is by no means an exhaustive survey of the literature on compressive sampling. Rather this is merely an account of the author's own work and thinking in this area which also includes a fairly large number of references to other people's work and occasionally discusses connections with these works. We have done our best to organize the ideas into a logical progression starting with the early papers which launched this subject. Before we begin, we would like to invite the interested reader to also check the article [17] by Ronald DeVore – also in these proceedings – for a complementary survey of the field (Section 5).

## 2. Undersampled measurements

Consider the general problem of reconstructing a vector $x \in \mathbb{R}^N$ from linear measurements $y$ about $x$ of the form

$$y_k = \langle x, \varphi_k \rangle, \ k = 1, \dots, K, \quad \text{or} \quad y = \Phi x. \tag{2.1}$$

That is, we acquire information about the unknown signal by sensing $x$ against $K$ vectors $\varphi_k \in \mathbb{R}^N$. We are interested in the "underdetermined" case $K \ll N$, where we have many fewer measurements than unknown signal values. Problems of this type arise in a countless number of applications. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the number of unknown pixels. In wideband radio frequency signal analysis, one may only be able to acquire a signal at a rate which is far lower than the Nyquist rate because of current limitations in Analog-to-Digital Converter technology. Finally, gene expression studies also provide examples of this kind. Here, one would like to infer the gene expression level of thousands of genes – that is, the dimension $N$ of the vector $x$ is in the thousands – from a low number of observations, typically in the tens.

At first glance, solving the underdertermined system of equations appears hopeless, as it is easy to make up examples for which it clearly cannot be done. But suppose now that the signal $x$ is *compressible*, meaning that it essentially depends on a number of degrees of freedom which is smaller than $N$. For instance, suppose our signal is sparse in the sense that it can be written either exactly or accurately as a superposition of a small number of vectors in some fixed basis. Then this premise radically changes the problem, making the search for solutions feasible. In fact, accurate and sometimes exact recovery is possible by solving a simple convex optimization problem.

**2.1. A nonlinear sampling theorem.** It might be best to consider a concrete example first. Suppose here that one collects an incomplete set of frequency samples of a discrete signal $x$ of length $N$. (To ease the exposition, we consider a model problem in one dimension. The theory extends easily to higher dimensions. For instance, we could be equally interested in the reconstruction of 2- or 3-dimensional objects from undersampled Fourier data.) The goal is to reconstruct the full signal $f$ given only $K$ samples in the Fourier domain

$$y_k = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x_t \, e^{-i2\pi\omega_k t/N}, \tag{2.2}$$

where the 'visible' frequencies $\omega_k$ are a subset $\Omega$ (of size $K$) of the set of all frequencies $\{0, \ldots, N-1\}$. Sensing an object by measuring selected frequency coefficients is the principle underlying Magnetic Resonance Imaging, and is common in many fields of science, including Astrophysics. In the language of the general problem (2.1), the sensing matrix $\Phi$ is obtained by sampling $K$ rows of the $N$ by $N$ discrete Fourier transform matrix.

We will say that a vector $x$ is *S-sparse* if its support $\{i : x_i \neq 0\}$ is of cardinality less or equal to $S$. Then Candès, Romberg and Tao [6] showed that one could almost always recover the signal $x$ exactly by solving the convex program[1] ($\|\tilde{x}\|_{\ell_1} := \sum_{i=1}^N |\tilde{x}_i|$)

$$(\mathrm{P}_1) \qquad \min_{\tilde{x}\in\mathbb{R}^N} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \Phi\tilde{x} = y. \tag{2.3}$$

[1] ($\mathrm{P}_1$) can even be recast as a linear program [3], [15].

**Theorem 2.1** ([6]). *Assume that $x$ is $S$-sparse and that we are given $K$ Fourier coefficients with frequencies selected uniformly at random. Suppose that the number of observations obeys*

$$K \geq C \cdot S \cdot \log N. \tag{2.4}$$

*Then minimizing $\ell_1$ reconstructs $x$ exactly with overwhelming probability. In details, if the constant $C$ is of the form $22(\delta + 1)$ in* (2.4)*, then the probability of success exceeds $1 - O(N^{-\delta})$.*

The first conclusion is that one suffers no information loss by measuring just about any set of $K$ frequency coefficients. The second is that the signal $x$ can be exactly recovered by minimizing a convex functional which does not assume any knowledge about the number of nonzero coordinates of $x$, their locations, and their amplitudes which we assume are all completely unknown a priori.

While this seems to be a great feat, one could still ask whether this is optimal, or whether one could do with even fewer samples. The answer is that in general, we cannot reconstruct $S$-sparse signals with fewer samples. There are examples for which the minimum number of samples needed for exact reconstruction by any method, no matter how intractable, must be about $S \log N$. Hence, the theorem is tight and $\ell_1$-minimization succeeds nearly as soon as there is any hope to succeed by any algorithm.

The reader is certainly familiar with the Nyquist/Shannon sampling theory and one can reformulate our result to establish simple connections. By reversing the roles of time and frequency in the above example, we can recast Theorem 1 as a new nonlinear sampling theorem. Suppose that a signal $x$ has support $\Omega$ in the frequency domain with $B = |\Omega|$. If $\Omega$ is a connected set, we can think of $B$ as the bandwidth of $x$. If in addition the set $\Omega$ is known, then the classical Nyquist/Shannon sampling theorem states that $x$ can be reconstructed perfectly from $B$ equally spaced samples in the time domain[2]. The reconstruction is simply a linear interpolation with a "sinc" kernel.

Now suppose that the set $\Omega$, still of size $B$, is unknown and not necessarily connected. In this situation, the Nyquist/Shannon theory is unhelpful – we can only assume that the connected frequency support is the entire domain suggesting that *all* $N$ time-domain samples are needed for exact reconstruction. However, Theorem 2.1 asserts that far fewer samples are necessary. Solving $(P_1)$ will recover $x$ perfectly from about $B \log N$ time samples. What is more, these samples do not have to be carefully chosen; almost any sample set of this size will work. Thus we have a nonlinear analog (described as such since the reconstruction procedure $(P_1)$ is nonlinear) to Nyquist/Shannon: we can reconstruct a signal with *arbitrary and unknown* frequency support of size $B$ from about $B \log N$ *arbitrarily chosen* samples in the time domain.

Finally, we would like to emphasize that our Fourier sampling theorem is only a special instance of much more general statements. As a matter of fact, the results

[2]For the sake of convenience, we make the assumption that the bandwidth $B$ divides the signal length $N$ evenly.

extend to a variety of other setups and higher dimensions. For instance, [6] shows how one can reconstruct a piecewise constant (one or two-dimensional) object from incomplete frequency samples provided that the number of jumps (discontinuities) obeys the condition above by minimizing other convex functionals such as the total variation.

**2.2. Background.** Now for some background. In the mid-eighties, Santosa and Symes [44] had suggested the minimization of $\ell_1$-norms to recover sparse spike trains, see also [25], [22] for early results. In the last four years or so, a series of papers [26], [27], [28], [29], [33], [30] explained why $\ell_1$ could recover sparse signals in some special setups. We note though that the results in this body of work are very different than the sampling theorem we just introduced. Finally, we would like to point out important connections with the literature of theoretical computer science. Inspired by [37], Gilbert and her colleagues have shown that one could recover an $S$-sparse signal with probability exceeding $1-\delta$ from $S \cdot \mathrm{poly}(\log N, \log \delta)$ frequency samples placed on special equispaced grids [32]. The algorithms they use are not based on optimization but rather on ideas from the theory of computer science such as isolation, and group testing. Other points of connection include situations in which the set of spikes are spread out in a somewhat even manner in the time domain [22], [51].

**2.3. Undersampling structured signals.** The previous example showed that the structural content of the signal allows a drastic "undersampling" of the Fourier transform while still retaining enough information for exact recovery. In other words, if one wanted to sense a sparse object by taking as few measurements as possible, then one would be well-advised to measure randomly selected frequency coefficients. In truth, this observation triggered a massive literature. To what extent can we recover a compressible signal from just a few measurements. What are good sensing mechanisms? Does all this extend to object that are perhaps not sparse but well-approximated by sparse signals? In the remainder of this paper, we will provide some answers to these fundamental questions.

## 3. The Mathematics of compressive sampling

**3.1. Sparsity and incoherence.** In all what follows, we will adopt an abstract and general point of view when discussing the recovery of a vector $x \in \mathbb{R}^N$. In practical instances, the vector $x$ may be the coefficients of a signal $f \in \mathbb{R}^N$ in an orthonormal basis $\Psi$

$$f(t) = \sum_{i=1}^{N} x_i \psi_i(t), \quad t = 1, \ldots, N. \tag{3.1}$$

For example, we might choose to expand the signal as a superposition of spikes (the canonical basis of $\mathbb{R}^N$), sinusoids, $B$-splines, wavelets [36], and so on. As a side

note, it is not important to restrict attention to orthogonal expansions as the theory and practice of compressive sampling accommodates other types of expansions. For example, $x$ might be the coefficients of a digital image in a tight-frame of curvelets [5]. To keep on using convenient matrix notations, one can write the decomposition (3.1) as $x = \Psi f$ where $\Psi$ is the $N$ by $N$ matrix with the waveforms $\psi_i$ as rows or equivalently, $f = \Psi^* x$.

We will say that a signal $f$ is sparse in the $\Psi$-domain if the coefficient sequence is supported on a small set and compressible if the sequence is concentrated near a small set. Suppose we have available undersampled data about $f$ of the same form as before

$$y = \Phi f.$$

Expressed in a different way, we collect partial information about $x$ via $y = \Phi' x$ where $\Phi' = \Phi\Psi^*$. In this setup, one would recover $f$ by finding – among all coefficient sequences consistent with the data – the decomposition with minimum $\ell_1$-norm

$$\min \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \Phi'\tilde{x} = y.$$

Of course, this is the same problem as (2.3), which justifies our abstract and general treatment.

With this in mind, the key concept underlying the theory of compressive sampling is a kind of uncertainty relation, which we explain next.

**3.2. Recovery of sparse signals.** In [7], Candès and Tao introduced the notion of uniform uncertainty principle (UUP) which they refined in [8]. The UUP essentially states that the $K \times N$ sensing matrix $\Phi$ obeys a "restricted isometry hypothesis." Let $\Phi_T,\ T \subset \{1, \ldots, N\}$ be the $K \times |T|$ submatrix obtained by extracting the columns of $\Phi$ corresponding to the indices in $T$; then [8] defines the $S$-restricted isometry constant $\delta_S$ of $\Phi$ which is the smallest quantity such that

$$(1 - \delta_S)\,\|c\|_{\ell_2}^2 \le \|\Phi_T c\|_{\ell_2}^2 \le (1 + \delta_S)\,\|c\|_{\ell_2}^2 \tag{3.2}$$

for all subsets $T$ with $|T| \le S$ and coefficient sequences $(c_j)_{j\in T}$. This property essentially requires that every set of columns with cardinality less than $S$ approximately behaves like an orthonormal system.

An important result is that if the columns of the sensing matrix $\Phi$ are approximately orthogonal, then the exact recovery phenomenon occurs.

**Theorem 3.1** ([8]). *Assume that $x$ is $S$-sparse and suppose that $\delta_{2S} + \delta_{3S} < 1$ or, better, $\delta_{2S} + \theta_{S,2S} < 1$. Then the solution $x^\star$ to* (2.3) *is exact, i.e., $x^\star = x$.*

In short, if the UUP holds at about the level $S$, the minimum $\ell_1$-norm reconstruction is provably exact. The first thing one should notice when comparing this result with the Fourier sampling theorem is that it is deterministic in the sense that it does not involve any probabilities. It is also universal in that *all* sufficiently sparse vectors

are exactly reconstructed from $\Phi x$. In Section 3.4, we shall give concrete examples of sensing matrices obeying the exact reconstruction property for large values of the sparsity level, e.g. for $S = O(K/\log(N/K))$.

Before we do so, however, we would like to comment on the slightly better version $\delta_{2S} + \theta_{S,2S} < 1$, which is established in [10]. The number $\theta_{S,S'}$ for $S + S' \leq N$ is called the $S, S'$-*restricted orthogonality constants* and is the smallest quantity such that

$$|\langle \Phi_T c, \Phi_{T'} c' \rangle| \leq \theta_{S,S'} \cdot \|c\|_{\ell_2} \|c'\|_{\ell_2} \tag{3.3}$$

holds for all *disjoint* sets $T, T' \subseteq \{1, \ldots, N\}$ of cardinality $|T| \leq S$ and $|T'| \leq S'$. Thus $\theta_{S,S'}$ is the cosine of the smallest angle between the two subspaces spanned by the columns in $T$ and $T'$. Small values of restricted orthogonality constants indicate that disjoint subsets of covariates span nearly orthogonal subspaces. The condition $\delta_{2S} + \theta_{S,2S} < 1$ is better than $\delta_{2S} + \delta_{3S} < 1$ since it is not hard to see that $\delta_{S+S'} - \delta_{S'} \leq \theta_{S,S'} \leq \delta_{S+S'}$ for $S' \geq S$ [8, Lemma 1.1].

Finally, now that we have introduced all the quantities needed to state our recovery theorem, we would like to elaborate on the condition $\delta_{2S} + \theta_{S,2S} < 1$. Suppose that $\delta_{2S} = 1$ which may indicate that there is a matrix $\Phi_{T_1 \cup T_2}$ with $2S$ columns ($|T_1| = S, |T_2| = S$) that is rank-deficient. If this is the case, then there is a pair $(x_1, x_2)$ of nonvanishing vectors with $x_1$ supported on $T_1$ and $x_2$ supported on $T_2$ obeying

$$\Phi(x_1 - x_2) = 0 \iff \Phi x_1 = \Phi x_2.$$

In other words, we have two very distinct $S$-sparse vectors which are indistinguishable. This is why any method whatsoever needs $\delta_{2S} < 1$. For, otherwise, the model is not identifiable to use a terminology borrowed from the statistics literature. With this in mind, one can see that the condition $\delta_{2S} + \theta_{S,2S} < 1$ is only slightly stronger than this identifiability condition.

**3.3. Recovery of compressible signals.** In general, signals of practical interest may not be supported in space or in a transform domain on a set of relatively small size. Instead, they may only be concentrated near a sparse set. For example, a commonly discussed model in mathematical image or signal processing assumes that the coefficients of elements taken from a signal class decay rapidly, typically like a power law. Smooth signals, piecewise signals, images with bounded variations or bounded Besov norms are all of this type [24].

A natural question is how well one can recover a signal that is just nearly sparse. For an arbitrary vector $x$ in $\mathbb{R}^N$, denote by $x_S$ its best $S$-sparse approximation; that is, $x_S$ is the approximation obtained by keeping the $S$ largest entries of $x$ and setting the others to zero. It turns out that if the sensing matrix obeys the uniform uncertainty principle at level $S$, then the recovery error is not much worse than $\|x - x_S\|_{\ell_2}$.

**Theorem 3.2** ([9]). *Assume that $x$ is $S$-sparse and suppose that $\delta_{3S} + \delta_{4S} < 2$. Then the solution $x^\star$ to* (2.3) *obeys*

$$\|x^* - x\|_{\ell_2} \le C \cdot \frac{\|x - x_S\|_{\ell_1}}{\sqrt{S}}. \tag{3.4}$$

*For reasonable values of $\delta_{4S}$, the constant in* (3.4) *is well behaved; e.g. $C \le 8.77$ for $\delta_{4S} = 1/5$. Suppose further that $\delta_S + 2\theta_{S,S} + \theta_{2S,S} < 1$, we also have*

$$\|x^* - x\|_{\ell_1} \le C \, \|x - x_S\|_{\ell_1}, \tag{3.5}$$

*for some positive constant $C$. Again, the constant in* (3.5) *is well behaved.*

Roughly speaking, the theorem says that minimizing $\ell_1$ recovers the $S$-largest entries of an $N$-dimensional unknown vector $x$ from $K$ measurements only. As a side remark, the $\ell_2$-stability result (3.4) appears explicitly in [9] while the '$\ell_1$ instance optimality' (3.5) is implicit in [7] although it is not stated explicitely. For example, it follows from Lemma 2.1 – whose hypothesis holds because of Lemma 2.2. in [8] – in that paper. Indeed, let $T$ be the set where $x$ takes on its $S$-largest values. Then Lemma 2.1 in [7] gives $\|x^* \cdot 1_{T^c}\|_{\ell_1} \le 4\|x - x_S\|_{\ell_1}$ and, therefore, $\|(x^* - x) \cdot 1_{T^c}\|_{\ell_1} \le 5\|x - x_S\|_{\ell_1}$. We conclude by observing that on $T$ we have

$$\|(x^* - x) \cdot 1_T\|_{\ell_1} \le \sqrt{S} \, \|(x^* - x) \cdot 1_T\|_{\ell_2} \le C \, \|x - x_S\|_{\ell_1},$$

where the last inequality follows from (3.4). For information, a more direct argument yields better constants.

To appreciate the content of Theorem 3.2, suppose that $x$ belongs to a weak-$\ell_p$ ball of radius $R$. This says that if we rearrange the entries of $x$ in decreasing order of magnitude $|x|_{(1)} \ge |x|_{(2)} \ge \cdots \ge |x|_{(N)}$, the $i$th largest entry obeys

$$|x|_{(i)} \le R \cdot i^{-1/p}, \quad 1 \le i \le N. \tag{3.6}$$

More prosaically, the coefficient sequence decays like a power-law and the parameter $p$ controls the speed of the decay: the smaller $p$, the faster the decay. Classical calculations then show that the best $S$-term approximation of an object $x \in w\ell_p(R)$ obeys

$$\|x - x_S\|_{\ell_2} \le C_2 \cdot R \cdot S^{1/2 - 1/p} \tag{3.7}$$

in the $\ell_2$ norm (for some positive constant $C_2$), and

$$\|x - x_S\|_{\ell_1} \le C_1 \cdot R \cdot S^{1 - 1/p}$$

in the $\ell_1$-norm. For generic elements obeying (3.6), there are no fundamentally better estimates available. Hence, Theorem 3.2 shows that with $K$ measurements only, we can achieve an approximation error which is as good as that one would obtain by knowing everything about the signal and selecting its $S$-largest entries.

**3.4. Random matrices.** Presumably all of this would be interesting if one could design a sensing matrix which would allow us to recover as many entries of $x$ as possible with as few as $K$ measurements. In the language of Theorem 3.1, we would like the condition $\delta_{2S} + \theta_{S,2S} < 1$ to hold for large values of $S$, ideally of the order of $K$. This poses a design problem. How should one design a matrix $\Phi$ – that is to say, a collection of $N$ vectors in $K$ dimensions – so that any subset of columns of size about $S$ be about orthogonal? And for what values of $S$ is this possible?

While it might be difficult to exhibit a matrix which provably obeys the UUP for very large values of $S$, we know that trivial randomized constructions will do so with overwhelming probability. We give an example. Sample $N$ vectors on the unit sphere of $\mathbb{R}^K$ independently and uniformly at random. Then the condition of Theorems 3.1 and 3.2 hold for $S = O(K/\log(N/K))$ with probability $1 - \pi_N$ where $\pi_N = O(e^{-\gamma N})$ for some $\gamma > 0$. The reason why this holds may be explained by some sort of "blessing of high-dimensionality." Because the high-dimensional sphere is mostly empty, it is possible to pack many vectors while maintaining approximate orthogonality.

- *Gaussian measurements.* Here we assume that the entries of the $K$ by $N$ sensing matrix $\Phi$ are independently sampled from the normal distribution with mean zero and variance $1/K$. Then if

$$S \leq C \cdot K/\log(N/K), \tag{3.8}$$

  $S$ obeys the condition of Theorems 3.1 and 3.2 with probability $1 - O(e^{-\gamma N})$ for some $\gamma > 0$. The proof uses known concentration results about the singular values of Gaussian matrices [16], [45].

- *Binary measurements.* Suppose that the entries of the $K$ by $N$ sensing matrix $\Phi$ are independently sampled from the symmetric Bernoulli distribution $P(\Phi_{ki} = \pm 1/\sqrt{K}) = 1/2$. Then it is conjectured that the conditions of Theorems 3.1 and 3.2 are satisfied with probability $1 - O(e^{-\gamma N})$ for some $\gamma > 0$ provided that $S$ obeys (3.8). The proof of this fact would probably follow from new concentration results about the smallest singular value of a subgaussian matrix [38]. Note that the exact reconstruction property for $S$-sparse signals and (3.7) with $S$ obeying (3.8) are known to hold for binary measurements [7].

- *Fourier measurements.* Suppose now that $\Phi$ is a partial Fourier matrix obtained by selecting $K$ rows uniformly at random as before, and renormalizing the columns so that they are unit-normed. Then Candès and Tao [7] showed that Theorem 3.1 holds with overwhelming probability if $S \leq C \cdot K/(\log N)^6$. Recently, Rudelson and Vershynin [43] improved this result and established $S \leq C \cdot K/(\log N)^4$. This result is nontrivial and use sophisticated techniques from geometric functional analysis and probability in Banach spaces. It is conjectured that $S \leq C \cdot K/\log N$ holds.

- *Incoherent measurements.* Suppose now that $\Phi$ is obtained by selecting $K$ rows uniformly at random from an $N$ by $N$ orthonormal matrix $U$ and renormalizing the columns so that they are unit-normed. As before, we could think of $U$ as the matrix $\Phi\Psi^*$ which maps the object from the $\Psi$ to the $\Phi$-domain. Then the arguments used in [7], [43] to prove that the UUP holds for incomplete Fourier matrices extend to this more general situation. In particular, Theorem 3.1 holds with overwhelming probability provided that

$$S \le C \cdot \frac{1}{\mu^2} \cdot \frac{K}{(\log N)^4}, \tag{3.9}$$

  where $\mu := \sqrt{N}\,\max_{i,j}|U_{i,j}|$ (observe that for the Fourier matrix, $\mu = 1$ which gives the result in the special case of the Fourier ensemble above). With $U = \Phi\Psi^*$,

$$\mu := \sqrt{N}\,\max_{i,j}|\langle\varphi_i, \psi_j\rangle| \tag{3.10}$$

  which is referred to as the mutual coherence between the measurement basis $\Phi$ and the sparsity basis $\Psi$ [27], [28]. The greater the incoherence of the measurement/sparsity pair $(\Phi, \Psi)$, the smaller the number of measurements needed.

In short, one can establish the UUP for a few interesting random ensembles and we expect that in the future, many more results of this type will become available.

**3.5. Optimality.** Before concluding this section, it is interesting to specialize our recovery theorems to selected measurement ensembles now that we have established the UUP for concrete values of $S$. Consider the Gaussian measurement ensemble in which the entries of $\Phi$ are i.i.d. $N(0, 1/K)$. Our results say that one can recover any $S$-sparse vector from a random projection of dimension about $O(S \cdot \log(N/S))$, see also [18]. Next, suppose that $x$ is taken from a weak-$\ell_p$ ball of radius $R$ for some $0 < p < 1$, or from the $\ell_1$-ball of radius $R$ for $p = 1$. Then we have shown that for all $x \in w\ell_p(R)$

$$\|x^\star - x\|_{\ell_2} \le C \cdot R \cdot (K/\log(N/K))^{-r}, \quad r = 1/p - 1/2, \tag{3.11}$$

which has also been proven in [20]. An important question is whether this is optimal. In other words, can we find a possibly adaptive set of measurements and a reconstruction algorithm that would yield a better bound than (3.11)? By adaptive, we mean that one could use a sequential measurement procedure where at each stage, one would have the option to decide which linear functional to use next based on the data collected up to that stage.

It proves to be the case that one cannot improve on (3.11), and we have thus identified the optimal performance. Fix a class of object $\mathcal{F}$ and let $E_K(\mathcal{F})$ be the best reconstruction error from $K$ linear measurements

$$E_K(\mathcal{F}) = \inf \sup_{f\in\mathcal{F}} \|f - D(y)\|_{\ell_2}, \quad y = \Phi f, \tag{3.12}$$

where the infimum is taken over all set of $K$ linear functionals and all reconstruction algorithms $D$. Then it turns out $E_K(\mathcal{F})$ nearly equals the *Gelfand* numbers of a class $\mathcal{F}$ defined as

$$d_K(\mathcal{F}) = \inf_V \{\sup_{f\in\mathcal{F}} \|P_V f\| : \operatorname{codim}(V) < K\}, \tag{3.13}$$

where $P_V$ is the orthonormal projection on the subspace $V$. Gelfand numbers play an important role in approximation theory, see [40] for more information. If $\mathcal{F} = -\mathcal{F}$ and $\mathcal{F} = \mathcal{F}+\mathcal{F} \le c_{\mathcal{F}}\mathcal{F}$, then $d_K(\mathcal{F}) \le E_K(\mathcal{F}) \le c_{\mathcal{F}} d_K(\mathcal{F})$. Note that $c_{\mathcal{F}} = 2^{1/p}$ in the case where $\mathcal{F}$ is a weak-$\ell_p$ ball. The thing is that we know the approximate values of the Gelfand numbers for many classes of interest. Suppose for example that $\mathcal{F}$ is the $\ell_1$-ball of radius $R$. A seminal result of Kashin [35] and improved by Garnaev and Gluskin [31] shows that for this ball, the Gelfand numbers obey

$$C_1 \cdot R \cdot \sqrt{\frac{\log(N/K)+1}{K}} \le d_k(\mathcal{F}) \le C_2 \cdot R \cdot \sqrt{\frac{\log(N/K)+1}{K}}, \tag{3.14}$$

where $C_1, C_2$ are universal constants. Gelfand numbers are also approximately known for weak-$\ell_p$ balls as well; the only difference is that $((\log(N/K)+1)/K)^r$ substitutes $((\log(N/K)+1)/K)^{1/2}$. Hence, Kashin, Garnaev and Gluskin assert that with $K$ measurements, the minimal reconstruction error (3.12) one can hope for is bounded below by a constant times $(K/\log(N/K))^{-r}$. Kashin's arguments [35] also used probabilistic functionals which establish the existence of recovery procedures for which the reconstruction error is bounded above by the right-hand side of (3.14). Similar types of recovery have also been known to be possible in the literature of theoretical computer science, at least in principle, for certain types of random measurements [1].

In this sense, our results – specialized to Gaussian measurements – are optimal for weak-$\ell_p$ norms. The novelty is that the information about the object can be retrieved from random coefficients by minimizing a simple linear program (2.3), and that the decoding algorithm adapts automatically to the weak-$\ell_p$ signal class, without knowledge thereof. Minimizing the $\ell_1$-norm is adaptive and nearly gives the best possible reconstruction error simultaneously over a wide range of sparse classes of signals; no information about $p$ and the radius $R$ are required.

## 4. Robust compressive sampling

In any realistic application, we cannot expect to measure $\Phi x$ without any error, and we now turn our attention to the robustness of compressive sampling vis a vis measurement errors. This is a very important issue because any real-world sensor is subject to at least a small amount of noise. And one thus immediately understands that to be widely applicable, the methodology needs to be stable. Small perturbations in the

observed data should induce small perturbations in the reconstruction. Fortunately, the recovery procedures may be adapted to be surprisingly stable and robust vis a vis arbitrary perturbations.

Suppose our observations are inaccurate and consider the model

$$y = \Phi x + e, \tag{4.1}$$

where $e$ is a stochastic or deterministic error term with bounded energy $\|e\|_{\ell_2} \le \varepsilon$. Because we have inaccurate measurements, we now use a noise-aware variant of (2.3) which relaxes the data fidelity term. We propose a reconstruction program of the form

$$(\mathrm{P}_2) \qquad \min \ \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|\Phi\tilde{x} - y\|_{\ell_2} \le \varepsilon. \tag{4.2}$$

The difference with $(\mathrm{P}_1)$ is that we only ask the reconstruction be consistent with the data in the sense that $y - \Phi x^\star$ be within the noise level. The program $(\mathrm{P}_2)$ has a unique solution, is again convex, and is a special instance of a second order cone program (SOCP) [4].

**Theorem 4.1** ([9]). *Suppose that $x$ is an arbitrary vector in $\mathbb{R}^N$. Under the hypothesis of Theorem* 3.2*, the solution $x^\star$ to* $(\mathrm{P}_2)$ *obeys*

$$\|x^\star - x\|_{\ell_2} \le C_{1,S} \cdot \varepsilon + C_{2,S} \cdot \frac{\|x_0 - x_{0,S}\|_{\ell_1}}{\sqrt{S}}. \tag{4.3}$$

*For reasonable values of $\delta_{4S}$ the constants in* (4.3) *are well behaved, see* [9].

We would like to offer two comments. The first is that the reconstruction error is finite. This quiet observation is noteworthy because we recall that the matrix $\Phi$ is rectangular with many more columns than rows – thus having a fraction of vanishing singular values. Having said that, the mere fact that the severely ill-posed matrix inversion keeps the perturbation from "blowing up" may seem a little unexpected. Next and upon closer inspection, one sees that the reconstruction error is the sum of two terms: the first is simply proportional to the size of the measurement error while the second is the approximation error one would obtain in the noiseless case. In other words, the performance of the reconstruction degrades gracefully as the measurement noise increases. This brings us to our second point. In fact, it is not difficult to see that no recovery method can perform fundamentally better for arbitrary perturbations of size $\varepsilon$ [9]. For related results for Gaussian sensing matrices, see [19].

## 5. Connections with statistical estimation

In the remainder of this paper, we shall briefly explore some connections with other fields, and we begin with statistics. Suppose now that the measurement errors in (4.1) are stochastic. More explicitly, suppose that the model is of the form

$$y = \Phi x + z, \tag{5.1}$$

where $z_1, \ldots, z_k$ are i.i.d. with mean zero and variance $\sigma^2$. In this section, we will assume that the $z_k$'s are Gaussian although nothing in our arguments heavily relies upon this assumption. The problem is again to recover $x$ from $y$ which is a central problem in statistics since this is just the classical multivariate linear regression problem. Because the practical environment has changed dramatically over the last two decades or so, applications have emerged in which the number of observations is small compared to the dimension of the object we wish to estimate – here, $K \leq N$. This new paradigm sometimes referred to as "high-dimensional data" is currently receiving much attention and, clearly, the emerging theory of compressive sampling might prove very relevant.

The results from the previous sections are directly applicable. Suppose that $x$ is $S$-sparse to simplify our exposition. Because $\|z\|_{\ell_2}^2$ is distributed as a chi-squared with $K$ degrees of freedom, the reconstruction (4.2) would obey

$$\|x^* - x\|_{\ell_2}^2 \leq C \cdot K\sigma^2 \tag{5.2}$$

with high probability. While this may seem acceptable to the nonspecialist, modern results in the literature suggest that one might be able to get a better accuracy. In particular, one would like an adaptive error bound which depends upon the complexity of the true unknown parameter vector $x \in \mathbb{R}^N$. For example, if $x$ only has $S$ significant coefficients, we would desire an error bound of size about $S\sigma^2$; the less complex the estimand, the smaller the squared-error loss. This poses an important question: can we design an estimator whose accuracy depends upon the information content of the object we wish to recover?

**5.1. Ideal model selection.** To get a sense of what is possible, consider regressing the data $y$ onto an arbitrary subset $T$ by the method of least squares. Define $\hat{x}[T]$ to be the least squares estimate whose restriction to the set $T$ is given by

$$\hat{x}_T[T] = (\Phi_T^T \Phi_T)^{-1} \Phi_T^T y, \tag{5.3}$$

and which vanishes outside $T$. Above, $\hat{x}_T[T]$ is the restriction of $\hat{x}[T]$ to $T$ and similarly for $x_T$. Since $\hat{x}[T]$ vanishes outside $T$, we have

$$\mathbf{E}\|x - \hat{x}[T]\|^2 = \|x_T - \hat{x}_T[T]\|^2 + \sum_{i \notin T} |x_i|^2,$$

Consider the first term. We have

$$x_T - \hat{x}_T[T] = (\Phi_T^T \Phi_T)^{-1} \Phi_T^T (s + z),$$

where $s = \Phi_{T^c} x_{T^c}$. It follows that

$$\mathbf{E}\|x_T - \hat{x}_T[T]\|^2 = \|(\Phi_T^T \Phi_T)^{-1} \Phi_T^T s\|^2 + \sigma^2 \operatorname{Tr}((\Phi_T^T \Phi_T)^{-1}).$$

However, since all the eigenvalues of $\Phi_T^T \Phi_T$ belong to the interval $[1-\delta_{|T|}, 1+\delta_{|T|}]$, we have

$$\mathbf{E}\|x_T - \hat{x}_T[T]\|^2 \geq \frac{1}{1+\delta_{|T|}} \cdot |T| \cdot \sigma^2.$$

For each set $T$ with $|T| \leq S$ and $\delta_S < 1$, we then have

$$\mathbf{E}\|x - \hat{x}[T]\|^2 \geq \sum_{i \in T^c c} x_i^2 + \frac{1}{2}\,|T| \cdot \sigma^2.$$

We now search for an *ideal estimator* which selects that estimator $\hat{x}[T^*]$ from the family $(\hat{x}[T])_{T \subset \{1,\ldots,N\}}$ with minimal Mean-Squared Error (MSE):

$$\hat{x}[T^*] = \mathrm{argmin}_{T \subset \{1,\ldots,N\}}\ \mathbf{E}\|x - \hat{x}[T]\|^2.$$

This estimator is ideal because we would of course not know which estimator $\hat{x}_T$ is best; that is, to achieve the ideal MSE, one would need an oracle that would tell us which model $T$ to choose.

We will consider this ideal estimator nevertheless and take its MSE as a benchmark. The ideal MSE is bounded below by

$$\begin{aligned}\mathbf{E}\|x - \hat{x}[T^*]\|^2 &\geq \frac{1}{2}\ \min_T \left(\|x - \hat{x}[T]\|^2 + |T| \cdot \sigma^2\right) \\ &= \frac{1}{2} \sum_i \min(x_i^2, \sigma^2). \end{aligned} \tag{5.4}$$

Letting $x_S$ be the best $S$-sparse approximation to $x$, another way to express the right-hand side (5.4) is in term of the classical trade-off between the approximation error and the number of terms being estimated times the noise level

$$\mathbf{E}\|x - \hat{x}_{T^*}\|^2 \geq \frac{1}{2} \inf_{S \geq 0} \left(\|x - x_S\|^2 + S\sigma^2\right).$$

Our question is of course whether there is a computationally efficient estimator which can mimic the ideal MSE.

**5.2. The Dantzig selector.** Assume for simplicity that the columns of $\Phi$ are normalized (there are straightforward variations to handle the general case). Then the Dantzig selector estimates $x$ by solving the convex program

$$\text{(DS)} \qquad \min_{\tilde{x} \in \mathbb{R}^N} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \sup_{1 \leq i \leq N} |(\Phi^T r)_i| \leq \lambda \cdot \sigma \tag{5.5}$$

for some $\lambda > 0$, and where $r$ is the vector of residuals

$$r = y - \Phi\tilde{x}. \tag{5.6}$$

The solution to this optimization problem is the minimum $\ell_1$-vector which is consistent with the observations. The constraints impose that the residual vector is within the noise level and does not correlate too well with the columns of $\Phi$. For information, there exist related, yet different proposals in the literature, and most notably the lasso introduced by [47], see also [15]. Again, the program (DS) is convex and can be recast as a linear program (LP).

The main result in this line of research is that the Dantzig selector is not only computationally tractable, it is also accurate.

**Theorem 5.1** ([10]). *Set* $\lambda := (1 + t^{-1})\sqrt{2\log p}$ *in* (5.5) *and suppose that* $x$ *is* $S$*-sparse with* $\delta_{2S} + \theta_{S,2S} < 1 - t$. *Then with very large probability, the Dantzig selector* $\hat{x}$ *solution to* (5.5) *obeys*

$$\|\hat{x} - x\|^2 \leq O(\log p) \cdot \Big(\sigma^2 + \sum_i \min(x_i^2, \sigma^2)\Big). \tag{5.7}$$

Our result says that the Dantzig selector achieves a loss within a logarithmic factor of the ideal mean squared error one would achieve with an *oracle* which would supply perfect information about which coordinates are nonzero, and which were above the noise level. To be complete, it is possible to obtain similar bounds on the MSE.

There are extensions of this result to signals which are not sparse but compressible, e.g. for signals which belong to weak-$\ell_p$ balls. What is interesting here is that in some instances, even though the number of measurements is much smaller than the dimension of the parameter vector $x$, the Dantzig selector recovers the minimax rate that one would get if we were able to measure all the coordinates of $x$ *directly* via $\tilde{y} = x + \sigma z$ where $z$ is i.i.d. $N(0, 1)$.

## 6. Connections with error correction

Compressive sampling also interacts with the agenda of coding theory. Imagine we wish to transmit a vector $x$ of length $M$ to a remote receiver reliably. A frequently discussed approach consists in encoding the information $x$ with an $N$ by $M$ coding matrix $C$ with $N > M$. Assume that gross errors occur upon transmission and that a fraction of the entries of $Cx$ are corrupted in a completely arbitrary fashion. We do not know which entries are affected nor do we know how they are affected. Is it possible to recover the information $x$ exactly from the corrupted $N$-dimensional vector $y$?

To decode, [8] proposes solving the minimum $\ell_1$-approximation problem

$$(\mathrm{D}_1) \qquad \min_{\tilde{x} \in \mathbb{R}^M} \|y - C\tilde{x}\|_{\ell_1}, \tag{6.1}$$

which can also be obviously recast as an LP. The result is that if $C$ is carefully chosen, then (6.1) will correctly retrieve the information $x$ with no error provided that the

fraction $\rho$ of errors is not too large, $\rho \leq \rho^*$. This phenomenon holds for all $x$'s and all corruption patterns.

To see why this phenomenon occurs, consider a matrix $B$ which annihilates the $N \times M$ coding matrix $C$ on the left, i.e. such that $BC = 0$; $B$ is called a parity-check matrix and is any $(N - M) \times N$ matrix whose kernel is the range of $C$ in $\mathbb{R}^N$. The transmitted information is of the form $y = Cx + e$, where $e$ is a sparse vector of possibly gross errors, and apply $B$ on both sides of this equation. This gives

$$\tilde{y} = B(Cx + e) = Be \tag{6.2}$$

since $BC = 0$. Therefore, the decoding problem is reduced to that of recovering the error vector $e$ from the observations $Be$. Once $e$ is known, $Cx$ is known and, therefore, $x$ is also known since we may just assume that $C$ has full rank.

Now the reader knows that we could solve the underdetermined system (6.2) by $\ell_1$-minimization. He also knows that if the UUP holds, the recovery is exact. Now $(\mathrm{D}_1)$ and $(\mathrm{P}_1)$ are equivalent programs. Indeed, it follows from the decomposition $\tilde{x} = x + h$ that

$$(\mathrm{D}_1) \iff \min_{h \in \mathbb{R}^M} \|e - Ch\|_{\ell_1}.$$

Now the constraint $Bd = Be$ means that $d = e - Ah$ for some $h \in \mathbb{R}^M$ and, therefore,

$$\begin{aligned} \min \|d\|_{\ell_1}, \quad Bd = Be &\iff \min_{h \in \mathbb{R}^n} \|d\|_{\ell_1}, \quad d = e - Ah \\ &\iff \min_{h \in \mathbb{R}^n} \|e - Ah\|_{\ell_1}, \end{aligned}$$

which proves the claim.

Hence, if one uses a random coding matrix which is a popular choice, we have the following result, see also [42]:

**Theorem 6.1** ([8]). *Suppose the coding matrix $C$ has i.i.d. $N(0, 1)$ entries. Then with probability exceeding $1 - O(e^{-\gamma M})$ for some $\gamma > 0$, $(\mathrm{D}_1)$ exactly decodes all $x \in \mathbb{R}^M$ provided that the fraction $\rho$ of arbitrary errors obeys $\rho \leq \rho^*(M, N)$.*

In conclusion, one can correct a constant fraction of errors with arbitrary magnitudes by solving a convenient LP. In [8], the authors reported on numerical results showing that in practice $(\mathrm{D}_1)$ works extremely well and recovers the vector $x$ exactly provided that the fraction of the corrupted entries be less than about 17% in the case where $N = 2M$ and less than about 34% in the case where $N = 4M$.

## 7. Further topics

Our intention in this short survey was merely to introduce the new compressive sampling concepts. We presented an approach based on the notion of uncertainty principle

which gives a powerful and unified treatment of some of the main results underlying this theory. As we have seen, the UUP gives conditions for exact, approximate, and stable recoveries which are almost necessary. Another advantage that one can hardly neglect is that this makes the exposition fairly simple. Having said that, the early papers on compressive sampling – e.g. [6], [7], [20] – have spurred a large and fascinating literature in which other approaches and ideas have been proposed. Rudelson and Vershynin have used tools from modern Banach space theory to derive powerful results for Gaussian ensembles [42], [14], [43]. In this area, Pajor and his colleagues have established the existence of abstract reconstruction procedures from subgaussian measurements (including random binary sensing matrices) with powerful reconstruction properties. In a different direction, Donoho and Tanner have leveraged results from polytope geometry to obtain very precise estimates about the minimal number of Gaussian measurements needed to reconstruct $S$-sparse signals [21], [23], see also [43]. Tropp and Gilbert reported results about the performance of greedy methods for compressive sampling [49]. Haupt and Nowak have quantified the performance of combinatorial optimization procedures for estimating a signal from undersampled random projections in noisy environments [34]. Finally, Rauhut has worked out variations on the Fourier sampling theorem in which a sparse continuous-time trigonometric polynomials is randomly sampled in time [41]. Because of space limitations, we are unfortunately unable to do complete justice to this rapidly growing literature.

We would like to emphasize that there are many aspects of compressive sampling that we have not touched. For example, we have not discussed the practical performance of this new theory. In fact, numerical experiments have shown that compressive sampling behaves extremely well in practice. For example, it has been shown that from $3S-4S$ nonadaptive measurements, one can reconstruct an approximation of an image in a fixed basis which is more precise than that one would get by measuring all the coefficients of the object in that basis and selecting the $S$ largest [13], [50]. Further, numerical simulations with noisy data show that compressive sampling is very stable and performs well in noisy environments. In practice, the constants appearing in Theorems 4.1 and 5.1 are very small, see [9] and [10] for empirical results.

We would like to close this article by returning to the main theme of this paper, which is that compressive sampling invites to rethink sensing mechanisms. Because if one were to collect a comparably small number of general linear measurements rather than the usual pixels, one could in principle reconstruct an image with essentially the same resolution as that one would obtain by measuring all the pixels. Therefore, if one could design incoherent sensors (i.e. measuring incoherent linear functionals), the payoff could be extremely large. Several teams have already reported progress in this direction. For example, a team led by Baraniuk and Kelly have proposed a new camera architecture that employs a digital micromirror array to perform optical calculations of linear projections of an image onto pseudorandom binary patterns [46], [52]. Compressive sampling may also address challenges in the processing of wideband radio frequency signals since high-speed analog-to-digital convertor

technology indicates that current capabilities fall well short of needs, and that hardware implementations of high precision Shannon-based conversion seem out of sight for decades to come. Finally, compressive sampling has already found applications in wireless sensor networks [2]. Here, compressive sampling allows of *energy efficient* estimation of sensor data with comparably few sensor nodes. The power of these estimation schemes is that they require no prior information about the sensed data. All these applications are novel and exciting. Others might just be around the corner.

## References


[1] Alon, N., Matias, Y., Szegedy, B., The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** (1999), 137–147.

[2] Bajwa, W. U., Haupt, J., Sayeed, A. M., Nowak, R., Compressive wireless sensing. In *Proc. 5th Intl. Conf. on Information Processing in Sensor Networks* (IPSN '06), Nashville, TN, 2006, 134–142.

[3] Bloomfield, P., Steiger, W., *Least Absolute Deviations: Theory, Applications, and Algorithms*. Progr. Probab. Statist. 6, Birkhäuser, Boston, MA, 1983.

[4] Boyd, S., Vandenberghe, L., *Convex Optimization*. Cambridge University Press, Cambridge 2004.

[5] Candès, E. J., Donoho, D. L. New tight frames of curvelets and optimal Representations of objects with piecewise $C^2$ singularities. *Comm. Pure Appl. Math.* **57** (2004), 219–266.

[6] Candès, E. J., Romberg, J., Tao, T., Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52** (2006), 489–509.

[7] Candès, E. J., Tao, T., Near-optimal signal recovery from random projections and universal encoding strategies. *IEEE Trans. Inform. Theory*, 2004, submitted.

[8] Candès, E. J., Tao, T., Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** (2005), 4203–4215.

[9] Candès, E. J., Romberg, J., Tao, T., Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** (8) (2005), 1207–1223.

[10] Candès, E. J., Tao, T., The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *Ann. Statist.*, to appear.

[11] Candès, E. J., Romberg, J., The role of sparsity and incoherence for exactly reconstructing a signal from limited measurements. Technical Report, California Institute of Technology, 2004.

[12] Candès, E. J., Romberg, J., Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.* **6** (2) (2006), 227–254.

[13] Candès, E. J., Romberg, J., Practical signal recovery from random projections. In *SPIE International Symposium on Electronic Imaging: Computational Imaging III*, San Jose, California, January 2005.

[14] Candès, E. J., Rudelson, M., Vershynin, R. and Tao, T. Error correction via linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2005), IEEE Comput. Soc. Press, LosAlamitos, CA,295–308.

[15] Chen, S. S., Donoho, D. L., Saunders, M. A, Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** (1999), 33–61.

[16] Davidson, K. R., Szarek, S. J., Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces* (ed. by W. B. Johnson, J. Lindenstrauss), Vol. I, North-Holland, Amsterdam 2001, 317–366; Corrigendum, Vol. 2, 2003, 1819–1820.

[17] DeVore, R. A., Optimal computation. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2006/2007.

[18] Donoho, D. L., For most large underdetermined systems of linear equations the minimal $\ell_1$-norm solution is also the sparsest Solution. *Comm. Pure Appl. Math.* **59** (2006), 797–829.

[19] Donoho, D. L., For most large underdetermined systems of equations, the minimal $\ell_1$-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** (2006), 907–934.

[20] Donoho, D. L., Compressed sensing. Technical Report, Stanford University, 2004.

[21] Donoho, D. L., Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical Report, Stanford University, 2005.

[22] Donoho, D. L., Logan, B. F., Signal recovery and the large sieve, *SIAM J. Appl. Math.* **52** (1992), 577–591.

[23] Donoho, D. L., Tanner, J., Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** (2005), 9452–9457.

[24] Donoho, D. L., Vetterli, M., DeVore, R. A., Daubechies, I., Data compression and harmonic analysis. *IEEE Trans. Inform. Theory* **44** (1998), 2435–2476.

[25] Donoho, D. L., Stark, P. B., Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* **49** (1989), 906–931.

[26] Donoho, D. L., Huo, X., Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), 2845–2862.

[27] Donoho, D. L., Elad, M., Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1$ minimization. *Proc. Natl. Acad. Sci. USA* **100** (2003), 2197–2202.

[28] Elad, M., Bruckstein, A. M., A generalized uncertainty principle and sparse representation in pairs of $\mathbb{R}^N$ bases. *IEEE Trans. Inform. Theory* **48** (2002), 2558–2567.

[29] Feuer, A., Nemirovski, A., On sparse representation in pairs of bases. *IEEE Trans. Inform. Theory* **49** (2003), 1579–1581.

[30] Fuchs, J. J., On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory* **50** (2004), 1341–1344.

[31] Garnaev, A., Gluskin, E., The widths of a Euclidean ball. *Dokl. Akad. Nauk. USSR* **277** (1984), 1048–1052; English transl. *Soviet Math. Dokl.* **30** (1984), 200–204.

[32] Gilbert, A. C., Muthukrishnan, S., Strauss, M., Improved time bounds for near-optimal sparse Fourier representation. In *Proceedings of SPIE 5914* (Wavelets XI), ed. by M. Papadakis, A. F. Laine, M. A. Unser, 2005.

[33] Gribonval, R., Nielsen, M., Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* **49** (2003), 3320–3325.

[34] Haupt, J., Nowak, R., Signal reconstruction from noisy random projections. *IEEE Trans. Inform. Theory*, submitted.

[35] Kashin, B., The widths of certain finite dimensional sets and classes of smooth functions, *Izvestia* **41** (1977), 334–351.

[36] Mallat, S., *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.

[37] Mansour, Y., Randomized interpolation and approximation of sparse polynomials. *SIAM J. Comput.* **24** (1995), 357–368.

[38] Litvak, A. E., Pajor, A., Rudelson, M., Tomczak-Jaegermann, N., Smallest singular value of random matrices and geometry of random polytopes. Manuscript, 2004.

[39] Mendelson, S., Pajor, A., Tomczak-Jaegermann, N., Reconstruction and subgaussian processes. *C. R. Math. Acad. Sci. Paris* **340** (2005), 885–888.

[40] Pinkus, A., *N-Widths in Approximation Theory*. Ergeb. Math. Grenzgeb. (3) 7, Springer-Verlag, Berlin 1985.

[41] Rauhut, H., Random sampling of sparse trigonometric polynomials. Preprint, 2005.

[42] Rudelson, M., Vershynin, R., Geometric approach to error-correcting codes and reconstruction of signals. *Internat. Math. Res. Notices* **2005** (64) (2005), 4019–4041.

[43] Rudelson, M., Vershynin, R., Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. Preprint, 2006.

[44] Santosa, F., Symes, W. W., Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.* **7** (1986), 1307–1330.

[45] Szarek, S. J., Condition numbers of random matrices. *J. Complexity* **7** (1991), 131–149.

[46] Takhar, D., Laska, J. N., Wakin, M., Duarte, M. F., Baron, D., Sarvotham, S., Kelly, K. F., Baraniuk, R. G., A new compressive imaging camera architecture using optical-domain compression. *IS&T/SPIE Computational Imaging IV*, San Jose, January 2006.

[47] Tibshirani, R., Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996), 267–288.

[48] Tropp, J. A., Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** (2006), 1030–1051.

[49] Tropp, J. A., Gilbert, A. C., Signal recovery from partial information via orthogonal matching pursuit. Preprint, University of Michigan, 2005.

[50] Tsaig, Y., Donoho, D. L., Extensions of compressed sensing. Technical report, Department of Statistics, Stanford University, 2004.

[51] Vetterli, M., Marziliano, P., Blu, T., Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.* **50** (2002), 1417–1428.

[52] Wakin, M., Laska, J. N., Duarte, M. F., Baron, D., Sarvotham, S., Takhar, D., Kelly, K. F., Baraniuk, R. G., Compressive imaging for video representation and coding. *Picture Coding Symposium*, special session on Next Generation Video Representation, Beijing, April 2006.

Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125, U.S.A.

E-mail: emmanuel@acm.caltech.edu

# Total variation based image denoising and restoration

Vicent Caselles*

**Abstract.** This paper is devoted to the total variation (TV) based approach to image denoising and restoration. The restored image minimizes total variation in the class of images which satisfy the contraints given by the image acquisition model. We compute some explicit solutions of the denoising model which explain some of the features observed in numerical experiments. We also comment on some alternatives recently proposed by Y. Meyer which lead to $u + v$ image decompositions. Finally we propose a total variation approach to image restoration, i.e., deconvolution and denoising, in which the image acquisition model is incorporated as a set of local constraints.

**Mathematics Subject Classification (2000).** Primary 68U10; Secondary 65K10, 65J20, 94A08.

**Keywords.** Image restoration, total variation, variational methods.

## 1. Introduction

We assume that the image acquisition system may be modelled by the following image formation model

$$z = h * u + n, \tag{1}$$

where $u\colon \mathbb{R}^2 \to \mathbb{R}$ denotes the ideal undistorted image, $h\colon \mathbb{R}^2 \to \mathbb{R}$ is a blurring kernel, $z$ is the observed image which is represented as a function $z\colon \mathbb{R}^2 \to \mathbb{R}$, and $n$ is an additive Gaussian white noise with zero mean and standard deviation $\sigma$.

Let us denote by $\Omega$ the interval $(0, N]^2$. As in most of works, in order to simplify this problem, we shall assume that the functions $h$ and $u$ are periodic of period $N$ in each direction. That amounts to neglecting some boundary effects. Therefore, we shall assume that $h, u$ are functions defined in $\Omega$ and, to fix ideas, we assume that $h, u \in L^2(\Omega)$. Our problem is to recover as much as possible of $u$, from our knowledge of the blurring kernel $h$, the statistics of the noise $n$, and the observed image $z$.

The problem of recovering $u$ from $z$ is ill-posed due to the ill-conditioning of the operator $Hu = h * u$. Several methods have been proposed to recover $u$. Most of them can be classified as regularization methods which may take into account statistical properties (Wiener filters), information theoretic properties ([19]), a priori

*The author acknowledges partial support by the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya and by PNPGC project, reference BFM2003-02125.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

geometric models ([30]) or the functional analytic behavior of the image given in terms of its wavelet coefficients ([20]).

The typical strategy to solve this ill-conditioning is regularization. Probably one of the first examples of regularization method [31] consists in choosing between all possible solutions of (1) the one which minimized the Sobolev (semi) norm of $u$

$$\int_\Omega |Du|^2\,dx. \tag{2}$$

Usually, the only information we know about the noise is statistical and limited to an estimate of its mean and its variance. In that case, the model equation (1) is incorporated as a set of constraints for (2): a first constraint corresponding to the assumption that the noise has zero mean, and a second one translating the fact that $\sigma$ is an upper bound of the standard deviation of $n$.

This formulation was an important step, but the results were not satisfactory, mainly due to the unability of the previous functional to resolve discontinuities (edges) and oscillatory textured patterns. The smoothness required by the Dirichlet integral (2) is too restrictive and information corresponding to high frequencies of $z$ is attenuated by it. Indeed, functions in $W^{1,2}(\Omega)$ (i.e., functions $u \in L^2(\Omega)$ such that $Du \in L^2(\Omega)$) cannot have discontinuities along rectifiable curves. These observations motivated the introduction of total variation in image restoration problems by L. Rudin, S. Osher and E. Fatemi in their work [30]. The a priori hypothesis is that functions of bounded variation (the BV model) ([5]) are a reasonable functional model for many problems in image processing, in particular, for restoration problems ([30]). Typically, functions of bounded variation have discontinuities along rectifiable curves, being continuous in some sense (in the measure theoretic sense) away from discontinuities. The discontinuities could be identified with edges. The ability of total variation regularization to recover edges is one of the main features which advocates for the use of this model but its ability to describe textures is less clear, even if some textures can be recovered, up to a certain scale of oscillation.

On the basis of the BV model, Rudin–Osher–Fatemi [30] proposed to solve the following constrained minimization problem

$$\begin{aligned} &\text{Minimize} \quad \int_\Omega |Du| \\ &\text{subject to} \quad \int_\Omega |h*u(x) - z(x)|^2\,dx \le \sigma^2|\Omega|. \end{aligned} \tag{3}$$

Notice that the image acquisition model (1) is only incorporated through a global constraint. Notice also that, assuming that $h*1 = 1$ (energy preservation), the constraint that $\int_\Omega h*u\,dx = \int_\Omega z(x)$ is automatically satisfied by its minima [17]. In practice, the above problem is solved via the following unconstrained minimization problem

$$\text{Minimize} \quad \int_\Omega |Du| + \frac{\lambda}{2}\int_\Omega |h*u - z|^2\,dx \tag{4}$$

where the parameter $\lambda$ is positive. Recall that we may interpret $\lambda^{-1}$ as a penalization parameter which controls the trade-off between the goodness of fit of the constraint and the smoothness term given by the total variation. In this formulation, a methodology is required for a correct choice of $\lambda$. The connections between (3) and (4) were studied by A. Chambolle and P. L. Lions in [17] where they proved that both problems are equivalent for some positive value of the Lagrange multiplier $\lambda$.

A particular and important case contained in the above formulation is the denoising problem which corresponds to the case where $h = \delta$, so that equation (1) is written as $z = u + n$ where $n$ is an additive Gaussian white noise of zero mean and variance $\sigma^2$. In this case, the unconstrained variational formulation (5) with $h = \delta$ is

$$\text{Minimize} \quad \int_\Omega |Du| + \frac{\lambda}{2} \int_\Omega |u - z|^2 \, dx, \tag{5}$$

and it has been the object of much theoretical and numerical research (see [7] for a survey). Even if this model represented a theoretical and practical progress in the denoising problem due to the introduction of BV functions as image models, the experimental analysis readily showed its main drawbacks. Between them, let us mention the staircasing effect (when denoising a smooth ramp plus noise, the staircase is an admissible result), the pixelization of the image at smooth regions and the loose of fine textured regions, to mention some of them. This can be summarized with the simple observation that the residuals $z - u$, where $u$ represents the solution of (5), do not look like noise. The theoretical analysis of the behavior of solutions of (5) has been the objects of several works [3], [12], [13], [27], [26] and will be developed in Section 2 by exhibiting explicit solutions for specially constructed functions $z$.

In spite of this, a second life in the interest of total variation based regularization was initiated after the proposal of $u + v$ models by Y. Meyer in [26]. The solution $u$ of (5) permits to obtain a decomposition of the data $z$ as a sum of two components $u + v$ where $v$ is supposed to contain the noise and textured parts of the image $z$, while $u$ contains the geometric sketch of the image $z$. As Meyer observed, the $L^2$ norm of the residual $v := z - u$ in (5) is not the right one to obtain a decomposition of $z$ in terms of geometry plus texture and he proposed to measure the size of the textured part $v$ in terms of a dual BV norm showing that some models of texture have a small dual BV norm: this will be the object of Section 3.

The restoration problem (which corresponds to the case of nontrivial kernel $h$) has also been the object of much interest due to its applications in many contexts, like satellite, astronomical or video images, to mention a few of them. In Section 4 we shall discuss a total variation based approach to the restoration model in which the image acquisition model is incorporated as a set of local constraints. Indeed, when incorporating (1) as a constraint in (3) we loose the local character of (1) and the restored image does not look satisfactory in textured and smooth regions at the same time. Thus, we propose to incorporate (1) by ensuring that the residuals $z - h * u$ have a variance bounded by $\sigma^2$ in a sufficiently large region around each pixel (the sampling process is incorporated in the model), the size of the region has to be sufficient in

order to estimate the variance of the noise. This gives a constrained formulation of the problem with as many Lagrange multipliers as pixels, and a solution is computed using Uzawa's method. Finally, in Section 5 we display some experiments on restoration of satellite images which illustrate the results that can be obtained with this method.

## 2. Explicit solutions of TV based denoising

The constrained formulation of the total variation denoising is given by (3) with $h = \delta$. Its unconstrained formulation is given by (5) where $\lambda > 0$ is a penalization parameter. Both problems are equivalent for a certain value of $\lambda$ [17]. Our purpose in this section is to exhibit some qualitative features of total variation denoising by constructing explicit solutions of (5). Those features are well known at the experimental level, and the results give a theoretical justification of these observations. Our solutions will exhibit the possibility to resolve discontinuities, but also the loss of contrast, and the regularization of corners (thus, the image is loosing structure). The staircasing effect was explained in [27].

The construction of explicit solutions of (5) is related to the computation of solutions of the eigenvalue problem for the 1-Laplacian operator.

$$-\operatorname{div}\left(\frac{Du}{|Du|}\right) = u. \tag{6}$$

We denote by $\mathrm{BV}(\mathbb{R}^N)$ the space of functions of bounded variation in $\mathbb{R}^N$. For definitions concerning bounded variation functions we refer to [5]. The solution of (6) is understood in the following sense ([6], [7], [13]).

**Definition 2.1.** We say that a function $u \in L^2(\mathbb{R}^N) \cap \mathrm{BV}(\mathbb{R}^N)$ is a solution of (6) in $\mathbb{R}^N$ if there exits a vector field $\xi \in L^\infty(\mathbb{R}^N; \mathbb{R}^N)$ with $\|\xi\|_\infty \leq 1$, such that $(\xi, Du) = |Du|$ and

$$-\operatorname{div}\xi = u \quad \text{in } \mathcal{D}'(\mathbb{R}^N).$$

If the vector field $\xi \in L^\infty(\mathbb{R}^N; \mathbb{R}^N)$ is such that $\operatorname{div}\xi \in L^2(\mathbb{R}^N)$ and $u \in \mathrm{BV}(\mathbb{R}^N)$, the expression $(\xi, Du)$ is a distribution defined by the formula

$$\langle(\xi, Dw), \varphi\rangle := -\int_{R^N} w\,\varphi \operatorname{div}\xi\,dx - \int_{R^N} w\,\xi\cdot\nabla\varphi\,dx \quad \text{for all } \varphi \in C_0^\infty(\mathbb{R}^N).$$

Then $(\xi, Du)$ is a Radon measure in $\mathbb{R}^N$ which coincides with $\xi\cdot\nabla u$ when $u \in L^2(\mathbb{R}^N) \cap W^{1,1}(\mathbb{R}^N)$ [11].

The following result is taken from [13] and it explains how can we derive from solutions of (6) data $z$ for which the solution of (5) is explicit.

**Proposition 2.2.** *Let $u_i \in \mathrm{BV}(\mathbb{R}^N)$ be such that $\inf(|u_i|, |u_j|) = 0$, $i, j \in \{1, \ldots, m\}$, $i \neq j$. Assume that $u_i$ and $\sum_{i=1}^m u_i$ are solutions of the eigenvalue problem* (6),

*$i \in \{1, \dots, m\}$. Let $b_i \in R$, $i = 1, \dots, m$, $z := \sum_{i=1}^m b_i u_i$, and $\lambda > 0$. Then the solution $u$ of the variational problem* (5) *is* $u := \sum_{i=1}^m \operatorname{sign}(b_i)(|b_i| - \lambda^{-1})^+ u_i$.

Assume that $m = 1$ and $\bar{u}$ is a solution of (6). If $0 < \lambda^{-1} \le b$, then $u := a\bar{u}$ with $a = b - \lambda^{-1}$ is a solution of (5) for the datum $z = b\bar{u}$. Indeed, $u$ satisfies the Euler–Lagrange equation of (5) which characterizes its unique solution:

$$z = b\bar{u} = a\bar{u} + \lambda^{-1}\bar{u} = a\bar{u} - \lambda^{-1}\operatorname{div}\left(\frac{D\bar{u}}{|D\bar{u}|}\right) = u - \lambda^{-1}\operatorname{div}\left(\frac{Du}{|Du|}\right).$$

If $\lambda^{-1} > b$, then $u = 0$ is the solution of (5). Indeed, in this case $\|\lambda z\|_{\mathrm{BV}^*} \le 1$ (the dual norm in $\mathrm{BV}(\mathbb{R}^N)^*$) and there is a vector field $\xi \in L^\infty(\mathbb{R}^N; \mathbb{R}^N)$ with $\|\xi\|_\infty \le 1$, such that $-\operatorname{div}\xi = \lambda z$. Thus, $u = 0$ satisfies the Euler–Lagrange equation of (5). The proof when $b \le 0$ is similar and we skip the details. This solution exhibits a loss of contrast of size $\min(\lambda^{-1}, |b|)$ when the datum is $z = b\bar{u}$.

Our next theorem gives a family of solutions of (6) and is taken from [12] (see also [4]).

**Theorem 2.3.** *Let $C_1, \dots, C_m$ be bounded convex subsets of $\mathbb{R}^2$ which are disjoint. Let $b_i > 0$, $i = 1, \dots, m$, $k \in \{1, \dots, m\}$. Then $v := -\sum_{i=1}^k b_i \chi_{C_i} + \sum_{i=k+1}^m b_i \chi_{C_i}$ is a solution of* (6) *if and only if the following conditions holds.*

(i) *The sets $C_i$, $i = 1, \dots, m$, are of class $C^{1,1}$.*

(ii) *$b_i = \frac{P(C_i)}{|C_i|}$ for any $i \in \{1, \dots, m\}$.*

(iii) *The following inequalities hold:*

$$\operatorname*{ess\,sup}_{p \in \partial C_i} \kappa_{C_i}(p) \le \frac{P(C_i)}{|C_i|} \quad \text{for all } i = 1, \dots, m.$$

(iv) *If $E_1$ is a solution of the variational problem*

$$\min\left\{P(E) : \bigcup_{j=k+1}^m C_j \subseteq E \subseteq \mathbb{R}^2 \setminus \bigcup_{i=1}^k C_i\right\},$$

*then we have*

$$P(E_1) = \sum_{j=k+1}^m P(C_j).$$

*If $E_2$ is a solution of the variational problem*

$$\min\left\{P(E) : \bigcup_{i=1}^k C_i \subseteq E \subseteq \mathbb{R}^2 \setminus \bigcup_{j=k+1}^m C_j\right\},$$

*then we have*

$$P(E_2) = \sum_{i=1}^k P(C_i).$$

*Moreover, if $k = m$, then we do not need to assume that the $C_i$ are convex and we can replace condition* (i) *by the following one:*

(i′) *The sets $C_i$, $i = 1, \dots, m$, are convex and of class $C^{1,1}$.*

This result was essentially proved in [12] (though we only stated the result in its second assertion). Its extension to $\mathbb{R}^N$ was proved in [4] (replacing the curvature of the boundaries by the sum of principal curvatures) under the assumption that the sets $C_i$ are convex and of class $C^{1,1}$. Let us point out the following corollary for connected sets.

**Corollary 2.4.** *Let $C \subset \mathbb{R}^2$ be a bounded set of finite perimeter, and assume that $C$ is connected. The function $v := \lambda \chi_C$ is a solution of* (6) *if and only if the following three conditions hold.*

(i) $\lambda = \lambda_C := \frac{P(C)}{|C|}$.

(ii) *$C$ is convex and $\partial C$ is of class $C^{1,1}$.*

(iii) *The following inequality holds:*

$$\operatorname*{ess\,sup}_{p \in \partial C} \kappa_{\partial C}(p) \leq \frac{P(C)}{|C|}.$$

A convex set $C \subseteq \mathbb{R}^2$ such that $u := \lambda_C \chi_C$ is a solution of (6) is called calibrable. The above result gives a characterization of calibrable sets in $\mathbb{R}^2$ and was proved in [24], [12]. For convex sets in $\mathbb{R}^N$ of class $C^{1,1}$ the above result is true if we replace the curvature of the boundary by the sum of the principal curvatures [4].

**Example 1.** Let $C \subset \mathbb{R}^2$ be the set of Figure 1. It is easy to check that $C$ satisfies the

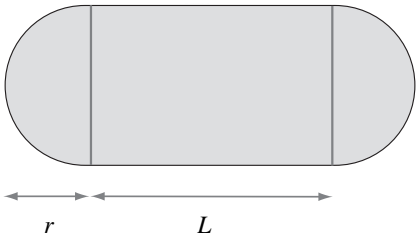



Figure 1. A bean-shaped set is calibrable.

assumptions of Corollary 2.4, since $C$ is a convex set with $C^{1,1}$ boundary and there holds

$$\operatorname*{ess\,sup}_{p \in \partial C} \kappa_{\partial C}(p) = \frac{1}{r} < \frac{2\pi r + 2L}{\pi r^2 + 2rL} = \frac{P(C)}{|C|}. \tag{7}$$

Moreover, since the inequality in (7) is always strict, any convex set $C'$ of class $C^{1,1}$ close enough to $C$ in the $C^{1,1}$-norm is also calibrable.

**Example 2.** Let $\Omega \subset \mathbb{R}^2$ be the union of two disjoint balls $B_1$ and $B_2$ of radius $r$, whose centers are at distance $L$ (see Figure 2). Then $k = 0$ and $m = 2$ in Theorem 2.3 and condition (iv) in it reads as

$$L \geq \pi r.$$

Under this condition the set $\Omega$ is calibrable. The condition $L \geq \pi r$ is nothing else than $P(\mathrm{co}(B_1 \cup B_2)) \geq P(B_1) + P(B_2)$ (co denotes the convex envelope) and in this case the solution of the denoising problem with $z = \chi_{B_1 \cup B_2}$ coincides with the

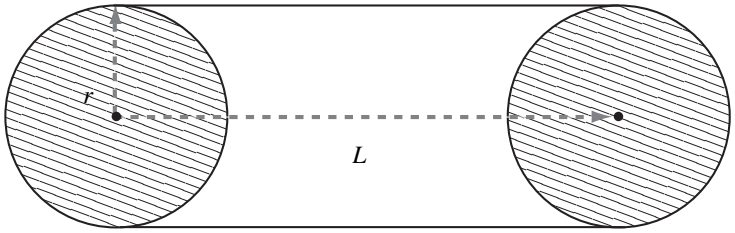



Figure 2. Two balls as initial datum for the denoising problem.

addition of the solutions obtained with $\chi_{B_1}$ and $\chi_{B_2}$. In case that $P(\mathrm{co}(B_1 \cup B_2)) < P(B_1) + P(B_2)$ there is interaction of the two sets and the solution is not the addition of solutions corresponding to the data $\chi_{B_1}$ and $\chi_{B_2}$.

These solutions exhibit two features of (5): that discontinuities may be preserved and the loss of contrast.

We could expand the above family of solutions by classifying all possible solutions of (6). Along this line, we extended the above results in two directions: in [13] we looked for solutions of (6) which are built up as sums of linear combinations of characteristic functions of convex sets of class $C^{1,1}$ (not disjoint, in general), and we considered in [3], [4] the case of general convex sets.

Let us illustrate the results in [13] with a simple case.

**Proposition 2.5.** *Let $K_0$, $K_1$ be two bounded open convex sets of $\mathbb{R}^2$ with boundary of class $C^{1,1}$ such that $\overline{K_1} \subseteq K_0$. Let $F := K_0 \setminus \overline{K_1}$. Let*

$$J := \frac{P(K_0) - P(K_1)}{|F|} > 0.$$

*If*

$$\operatorname*{ess\,sup}_{\partial K_0} \kappa_{\partial K_0} \leq J, \quad \operatorname*{ess\,inf}_{x \in \partial K_1} \kappa_{\partial K_1}(x) \geq J, \quad \operatorname*{ess\,sup}_{x \in \partial K_1} \kappa_{\partial K_1}(x) \leq \lambda_{K_1}$$

*then $v = \lambda_{K_1} \chi_{K_1} + J \chi_{K_0 \setminus K_1}$ is a solution of* (6).

The works [3], [4] describe the denoising of the characteristic function of any convex set of $\mathbb{R}^2$ and $\mathbb{R}^N$, respectively, and the results in them illustrate the regularization of corners. Even if the more general case of linear combinations of convex sets in $\mathbb{R}^2$ and $\mathbb{R}^N$ is considered, we illustrate the results in [3], [4] with a simple case.

**Theorem 2.6.** *Assume that $C$ is a bounded convex set in $\mathbb{R}^2$. Then there is a calibrable set $C_R \subseteq C$ such that $\partial C \setminus \partial C_R$ is formed by arcs of circle of radius $R$ such that $\frac{1}{R} = \frac{P(C_R)}{|C_R|}$ and for each $x \in C \setminus C_R$ it passes a unique arc of circle of radius $r(x)$ and those circles fiber $C \setminus C_R$. Let $r(x) = R$ for $x \in C_R$. Then $u(x) = \left(1 - \frac{\lambda^{-1}}{r(x)}\right)^+ \chi_C$ is the solution of* (6) *for the data $z = \chi_C$.*

## 3. Image decomposition models

In his work [26], Y. Meyer interpreted the denoising model as a $u + v$ decomposition. Assume that $\Omega$ is a bounded connected domain in $\mathbb{R}^2$ with Lipschitz boundary. If $z \in L^2(\Omega)$ and $u$ is the solution of (5), then its Euler–Lagrange equation can be written as

$$u + v = z \quad \text{where } v = -\frac{1}{\lambda}\operatorname{div}\left(\frac{Du}{|Du|}\right).$$

This type of decompositions is called a $u + v$ decomposition and $u$ is supposed to be a geometric sketch of the image [26]. As we have shown in the previous section, model (5) does not attain its objective of separating an image into its $u + v$ decomposition. This conclusion was also derived in [26] through complementary arguments. For instance, if $z = \chi_\omega$ where $\omega$ is a bounded domain with a $C^\infty$ boundary, then $z$ is not preserved by the Rudin–Osher–Fatemi (ROF) model (contrary to what it should be expected). The $v$ component contains the noise but also part of the image structure and, in particular, part of the texture (depending on the value of $\lambda$). On the other hand if $z(x) = \chi_A(x) + p(mx)\chi_B(x)$ where $A$ and $B$ are two bounded domains with smooth boundary, $m \geq 1$, and $p(x)$, $x = (x_1, x_2)$, is a smooth $2\pi$-periodic function of the two variables $x_1$, $x_2$, then the ROF model does not give $u(x) = \chi_A(x)$, $v(x) = p(mx)\chi_B(x)$ [26] (this will be explained after Theorem 3.1). Then to improve the ROF model Meyer proposed a different decomposition [26], which is based in the following variational model

$$\inf_{u \in \mathrm{BV}(\Omega), v \in G(\Omega), z = u+v} \int_\Omega |Du|\,dx + \lambda \|v\|_G,$$

where $\lambda > 0$ and $G(\Omega)$ denotes the Banach space of distributions $f$ in $\Omega$ that may be written

$$f = \operatorname{div} \xi$$

where $\xi \in L^\infty(\Omega; \mathbb{R}^2)$. The norm in $G$ is defined by

$$\|f\|_G := \inf\{\|\xi\|_\infty : \xi \in L^\infty(\Omega; \mathbb{R}^2),\ f = \operatorname{div}\xi\}$$

where $\|\xi\|_\infty := \operatorname{ess\,sup}_{x\in\Omega} |\xi(x)|$. $G(\Omega)$ is exactly $W^{-1,\infty}(\Omega)$, the dual space of $W_0^{1,1}(\Omega)$. The justification for the introduction of the space $G$ comes from the next result [26].

**Theorem 3.1.** *Let $f_n$ be a sequence of functions in $L^2(\Omega)$ with the following properties*

(i) *There exists a compact set $K \subset \Omega$ such that the support of $f_n$ is contained in $K$ for each $n$,*

(ii) *There exists $q > 2$ and $C > 0$ such that $\|f_n\|_q \leq C$,*

(iii) *The sequence $f_n$ converges to 0 in a distributional sense.*

*Then $\|f_n\|_G$ converges to 0 as $n \to \infty$.*

In other words, oscillating textures have a small norm in $G(\Omega)$. Now, if $z(x) = \chi_A(x) + p(mx)\chi_B(x)$ is as in the first paragraph of this section, then $v$ cannot be $p(mx)\chi_B(x)$ for large $m$ [26]. Otherwise we would have $p(mx)\chi_B(x) = -\frac{1}{\lambda}\operatorname{div}\left(\frac{Du}{|Du|}\right)$ and therefore $\|p(mx)\chi_B(x)\|_G = \frac{1}{\lambda}$. But we know from Theorem 3.1 that the $G$-norm of $p(mx)\chi_B(x)$ is small for large values $m$ (indeed the $G$-norm of $p(mx)\chi_B(x)$ is an $O(m^{-1})$ [26]).

Theorem 3.1 and other results [26], [25] were the starting point of extensive numerical work on $u + v$ decompositions [32], [28], [10], [9] to explore and compare the relative ability of the $G$ based model versus the ROF model. Meyer's model was first implemented by Vese–Osher in [32]. A different approach was proposed in [10], [9] where the decomposition is computed by minimizing a convex functional which depends on the two variable $u$ and $v$, alternatively in each variable. Each minimization is based on the projection algorithm introduced in [16]. The problem to solve is:

$$\inf_{(u,v)\in \mathrm{BV}(\Omega)\times\mu B_G} \int_\Omega |Du| + \frac{\lambda}{2}\int_\Omega |z-u-v|^2\,dx, \tag{8}$$

where $B_G := \{v \in G : \|v\|_G \leq 1\}$. We refer to [10] for its precise connection with Meyer's model. Let us mention that other dual Sobolev norms, indeed $H^{-1}$, have been explored in [28].

Figure 3 displays the comparison between ROF and model (8) for a simple figure. These images are courtesy of J. F. Aujol and A. Chambolle and have been obtained with the numerical methods developed by the authors in [9], [10]. Figures 3.a and 3.b display the original reference image and the noisy image with an additive Gaussian white noise with $\sigma = 35$. Figures 3.c and 3.d display the $u$ and $v$ components obtained with the ROF model with $\lambda$ chosen so that $\|v\| = \sigma$. For better visualization, the $v$ component will be always displayed as $v + 128$. Figures 3.e and 3.f display the $u$ and $v$ components obtained with model (8) with $\lambda = 10$ and $\mu = 55$ (for more details on the choice of parameters, see [9], [10]). In this case, for well chosen values of the parameter, the results are quite comparable. But let us point out that model (8) is able

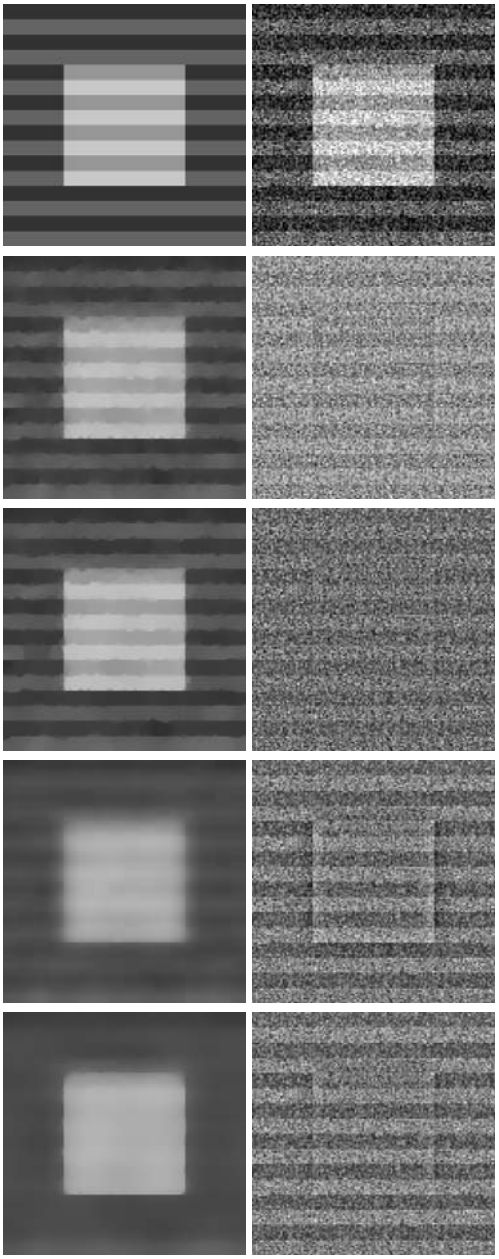


Figure 3. *Comparison of ROF and model* (8). From left to right and top to bottom: a) Original reference image. b) Noisy image with $\sigma = 35$. c) and d) Result of the ROF model: $u$ and $v$ component (with $\lambda$ chosen so that $\|v\| = \sigma$). For a better visualization, the $v$ component will be displayed as $v + 128$. e) and f) Result of model (8): $u$ and $v$ component ($\lambda = 10$, $\mu = 55$). g) and h) Result of the ROF model: $u$ and $v$ component (in this case $\sigma = 40.8$). i) and j) Result of model (8): $u$ and $v$ component (in this case $\sigma = 40.8$ and $\mu = 200$). These images are courtesy of J. F. Aujol and A. Chambolle. See the text for more details.

to separate the horizontal bands from the square for large values of the parameter $\mu$ while this does not seem to be possible with the ROF model. This is displayed in the next figures. Figures 3.g and 3.h display the $u$ and $v$ components obtained with the ROF model (the noise corresponds to a value of $\sigma = 40.8$) with $\lambda$ chosen so that $\|v\| = \sigma$. Figures 3.i and 3.j display the $u$ and $v$ components obtained with model (8) with $\mu = 200$. In any case, the choice of the parameters is open to further analysis and it the separation of the image in two components is related to the different scales present in the image.

## 4. Image restoration

To approach the problem of image restoration from a numerical point of view we shall assume that the image formation model incorporates the sampling process in a regular grid

$$z(i, j) = h * u(i, j) + n(i, j), \quad (i, j) \in \{1, \dots, N\}^2 \tag{9}$$

where $u\colon \mathbb{R}^2 \to \mathbb{R}$ denotes the ideal undistorted image, $h\colon \mathbb{R}^2 \to \mathbb{R}$ is a blurring kernel, $z$ is the observed sampled image which is represented as a function $z\colon \{1, \dots, N\}^2 \to \mathbb{R}$, and $n(i, j)$ is, as usual, a white Gaussian noise with zero mean and standard deviation $\sigma$.

Let us denote by $\Omega_N$ the interval $(0, N]^2$. As we said in the introduction, in order to simplify this problem, we assume that the functions $h$ and $u$ are periodic of period $N$ in each direction. That amounts to neglecting some boundary effects. Therefore, we assume that $h, u$ are functions defined in $\Omega_N$. To fix ideas, we assume that $h, u \in L^2(\Omega_N)$, so that $h * u$ is a continuous function in $\Omega_N$ (which may be extended to a continuous periodic function in $\mathbb{R}^2$) and the samples $h * u(i, j)$, $(i, j) \in \{1, \dots, N\}^2$, have sense.

Our next purpose is to introduce a restoration model with local constraints and to explain the numerical approach to solve it. For that, let us introduce some notation. We denote by $X$ the Euclidean space $\mathbb{R}^{N\times N}$. Then the image $u \in X$ is the vector $u = (u(i, j))_{i,j=1}^N$, and the vector field $\xi$ is the map $\xi\colon \{1, \dots, N\} \times \{1, \dots, N\} \to \mathbb{R}^2$. If $u \in X$, the discrete gradient is a vector in $Y = X \times X$ given by

$$\nabla^{+,+}u := (\nabla_x^+ u, \nabla_y^+ u),$$

where

$$\nabla_x^+ u(i, j) = \begin{cases} u(i+1, j) - u(i, j) & \text{if } i < N, \\ 0 & \text{if } i = N, \end{cases}$$

$$\nabla_y^+ u(i, j) = \begin{cases} u(i, j+1) - u(i, j) & \text{if } j < N, \\ 0 & \text{if } j = N, \end{cases}$$

for $i, j \in \{1, \dots, N\}$. We denote $\nabla^{+,+}u = (\nabla_x^+ u, \nabla_y^+ u)$. Other choices of the gradient are possible, this one will be convenient for the developments below.

Let us define the discrete functional

$$J_d^\beta(u) = \sum_{1 \le i,j \le N} \sqrt{\beta^2 + |\nabla^{+,+}u(i,j)|^2}, \quad \beta \ge 0.$$

For any function $w \in L^2(\Omega_N)$, its Fourier coefficients are

$$\hat{w}_{\frac{l}{N},\frac{j}{N}} = \int_{\Omega_N} w(x,y) e^{-2\pi i \frac{(lx+jy)}{N}} \quad \text{for } (l,j) \in \mathbb{Z}^2.$$

Our plan is to compute a band limited approximation to the solution of the restoration problem for (9). For that we define

$$\mathcal{B} := \left\{ u \in L^2(\Omega_N) : \hat{u} \text{ is supported in } \{-\tfrac{1}{2} + \tfrac{1}{N}, \dots, \tfrac{1}{2}\} \right\}.$$

We notice that $\mathcal{B}$ is a finite dimensional vector space of dimension $N^2$ which can be identified with $X$. Both $J(u) = \int_{\Omega_N} |Du|$ and $J_d^0(u)$ are norms on the quotient space $\mathcal{B}/\mathbb{R}$, hence they are equivalent. With a slight abuse of notation we shall indistinctly write $u \in \mathcal{B}$ or $u \in X$.

We shall assume that the convolution kernel $h \in L^2(\Omega_N)$ is such that $\hat{u}$ is supported in $\left\{ -\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2} \right\}$ and $\hat{h}(0,0) = 1$.

In the discrete framework, the ROF model for restoration is

$$\text{Minimize}_{u \in X} \quad J_d^\beta(u) \tag{10}$$

$$\text{subject to} \quad \sum_{i,j=1}^{N} |h * u(i,j) - z(i,j)|^2 \le \sigma^2 N^2. \tag{11}$$

Notice again that the image acquisition model (1) is only incorporated through a global constraint. In practice, the above problem is solved via the following unconstrained formulation

$$\min_{u \in X} \max_{\lambda \ge 0} J_d^\beta(u) + \frac{\lambda}{2}\Big[ \frac{1}{N^2} \sum_{i,j=1}^{N} |h * u(i,j) - z(i,j)|^2 - \sigma^2 \Big] \tag{12}$$

where $\lambda \ge 0$ is a Lagrange multiplier. The appropriate value of $\lambda$ can be computed using Uzawa's algorithm [15], [2] so that the constraint (11) is satisfied. Recall that if we interpret $\lambda^{-1}$ as a penalization parameter which controls the importance of the regularization term, and we set this parameter to be small, then homogeneous zones are well denoised while highly textured regions will loose a great part of its structure. On the contrary, if $\lambda^{-1}$ is set to be small, texture will be kept but noise will remain in homogeneous regions. On the other hand, as the authors of [15], [2] observed, if we use the constrained formulation (10)-(11) or, equivalently (12), then

the Lagrange multiplier does not produce satisfactory results since we do not keep textures and denoise flat regions simultaneously, and they proposed to incorporate the image acquisition model as a set of local constraints.

Following [2], we propose to replace the constraint (11) by

$$G*(h*u-z)(i,j) \le \sigma^2, \quad \text{for all } (i,j) \in \{1,\dots,N\}, \tag{13}$$

where $G$ is a discrete convolution kernel such that $G(i,j) > 0$ for all $(i,j) \in \{1,\dots,N\}$. The effective support of $G$ must permit the statistical estimation of the variance of the noise with (13) (see [2]). Then we shall minimize the functional $J_d^\beta(u)$ on $X$ submitted to the family of constraints (13) (plus eventually the constraint $\sum_{i,j=1}^N h*u(i,j) = \sum_{i,j=1}^N z(i,j)$). Thus, we propose to solve the optimization problem:

$$\begin{aligned} &\min_{u\in\mathcal{B}} J_d^\beta(u) \\ &\text{subject to} \quad G*(h*u-z)^2(i,j) \le \sigma^2 \quad \text{for all } (i,j). \end{aligned} \tag{14}$$

This problem is well-posed, i.e., there exists a solution and is unique if $\beta > 0$ and $\inf_{c\in\mathbb{R}} G*(z-c)^2 > \sigma^2$. In case that $\beta = 0$ and $\inf_{c\in\mathbb{R}} G*(z-c)^2 > \sigma^2$, then $h*u$ is unique. Moreover, it can be solved with a gradient descent approach and Uzawa's method [2].

To guarantee that the assumptions of Uzawa's method hold we shall use a gradient descent strategy. For that, let $v \in X$ and $\gamma > 0$. At each step we have to solve a problem like

$$\begin{aligned} &\min_{u\in\mathcal{B}} \gamma|u-v|_X^2 + J_d^\beta(u) \\ &\text{subject to} \quad G*(h*u-z)^2(i,j) \le \sigma^2 \quad \text{for all } (i,j). \end{aligned} \tag{15}$$

We solve (15) using the unconstrained formulation

$$\min_{u\in X} \max_{\lambda\ge 0} \mathcal{L}^\gamma(u,\{\lambda\};v),$$

where $\lambda = (\lambda(i,j))_{i,j=1}^N$ and

$$\mathcal{L}^\gamma(u,\{\lambda\};v) = \gamma|u-v|_X^2 + J_d^\beta(u) + \sum_{i,j=1}^N \lambda(i,j)(G*(h*u-z)^2(i,j)-\sigma^2).$$

**Algorithm: TV based restoration algorithm with local constraints**

1. Set $u^0 = 0$ or, better, $u^0 = z$. Set $n = 0$.
2. Use Uzawa's algorithm to solve the problem

$$\min_{u\in X} \max_{\lambda\ge 0} \mathcal{L}^\gamma(u,\{\lambda\};u^n), \tag{16}$$

that is:

(a) Choose any set of values $\lambda^0(i,\, j) \geq 0$, $(i,\, j) \in \{1, \ldots, N\}^2$, and $u_0^n = u^n$. Iterate from $p = 0$ until convergence of $\lambda^p$ the following steps:

(b) With the values of $\lambda^p$ solve the problem

$$\min_u \mathcal{L}^\gamma(u, \{\lambda^p\}; u^n)$$

starting with the initial condition $u_p^n$. Let $u_{p+1}^n$ be the solution obtained.

(c) Update $\lambda$ in the following way:

$$\lambda^{p+1}(i,\, j) = \max(\lambda^p(i,\, j) + \rho(G * (h * u_p^n - z)^2(i,\, j) - \sigma^2), 0)$$

for all $(i,\, j)$.

Let $u^{n+1}$ be the solution of (16). Stop when convergence of $u^n$.

We notice that, since $\gamma > 0$, Uzawa's algorithm converges if $z \in h * \mathcal{B}$. Moreover, if $u^0$ satisfies the constraints, then $u^n$ tends to a solution $u$ of (14) as $n \to \infty$ [2].

Finally, to solve problem (16) in Step 2.(b) of the algorithm we use either the extension of Chambolle's algorithm [16] to the restoration case given in [1] if we use $\beta = 0$, or the Bermúdez–Moreno algorithm [14] adapted to solve (16) when $\beta > 0$ as given in [2]. Being differentiable at when $\nabla^{+,+}u = 0$, this second possibility produces slightly smoother solutions in smooth non textured areas. We shall not enter on the comparison of both possibilities here and we shall use $\beta = 0$. For more details, we refer to [1], [2].

Let us mention the work [23] where the authors introduce a spatially varying fidelity term which controls the amount of denoising in any region of the image in order to preserve textures and small details. The philosophy is the same as ours but the value of $\lambda(i,\, j)$ is chosen in a different way.

## 5. Some restoration experiments

To simulate our data we use the modulation transfer function corresponding to SPOT 5 HRG satellite with Hipermode sampling (see [29] for more details):

$$\hat{h}(\eta_1, \eta_2) = e^{-4\pi\beta_1|\eta_1|} e^{-4\pi\alpha\sqrt{\eta_1^2+\eta_2^2}} \operatorname{sinc}(2\eta_1) \operatorname{sinc}(2\eta_2) \operatorname{sinc}(\eta_1), \qquad (17)$$

where $\eta_1, \eta_2 \in [-1/2, 1/2]$, $\operatorname{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_2)$, $\alpha = 0.58$, and $\beta_1 = 0.14$. Then we filter the reference image given in Figure 4.a with the filter (17) and we add some Gaussian white noise of zero mean and standard deviation $\sigma$ (in our case $\sigma = 1$, which is a realistic assumption for the case of satellite images [29]) to obtain the image displayed in Figure 4.b.

Figure 5.a displays the restoration of the image in Figure 4.b obtained using the algorithm of last section with $\beta = 0$. We have used a Gaussian function $G$ of standard deviation $\sigma = 6$. The mean value of the constraint is $\text{mean}((G*(Ku-z))^2) = 1.0933$ and RMSE $= 7.9862$. Figure 5.b displays the function $\lambda(i,\, j)$ obtained.

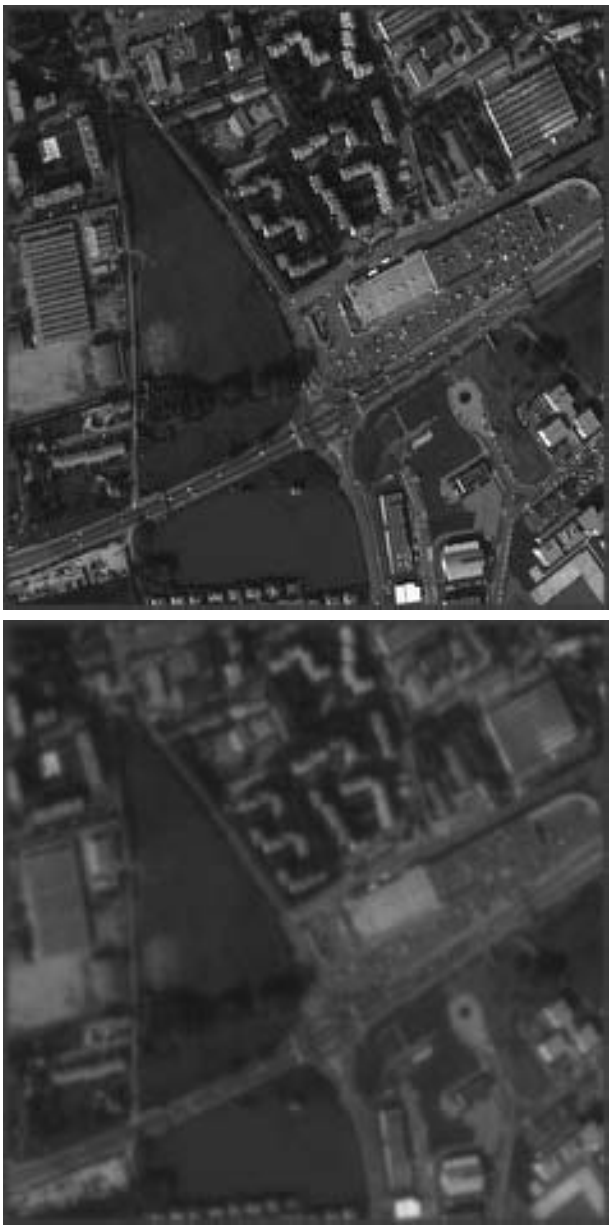


Figure 4. *Reference image and a filtered and noised image.* a) Top: reference image. b) Bottom: the data. This image has been generated applying the MTF given in (17) to the top image and adding a Gaussian white noise of zero mean and standard deviation $\sigma = 1$.

Figure 6 displays some details of the results that are obtained using a single global constraint (11) and show its main drawbacks. Figure 6.a corresponds to the result obtained with the Lagrange multiplier $\lambda = 10$ (thus, the constraint (11) is satisfied). The result is not satisfactory because it is difficult to denoise smooth regions and keep the textures at the same time. Figure 6.b shows that most textures are lost when using a small value of $\lambda$ ($\lambda = 2$) and Figure 6.c shows that some noise is present if we use a larger value of $\lambda$ ($\lambda = 1000$). This result is to be compared with the same detail of Figure 5.a which is displayed in Figure 6.d.

**The modulation transfer function for satellite images.** We describe here a simple model for the Modulation Transfer Function of a general satellite. More details can be found in [29] where specific examples of MTF for different acquisition systems are shown. The MTF used in our experiments (17) corresponds to a particular case of the general model described below [29].

Recall that the MTF, that we denote by $\hat{h}$, is the Fourier transform of the impulse response of the system. Let $(\eta_1, \eta_2) \in [-1/2, 1/2]$ denote the coordinates in the frequency domain. There are different parts in the acquisition system that contribute to the global transfer function:

**Sensors.** Every sensor has a sensitive region where all the photons that arrive are integrated. This region can be approximated by a unit square $[-c/2, c/2]^2$ where $c$ is the distance between consecutive sensors. Its impulse response is then the convolution of two pulses, one in each spatial direction. The corresponding transfer function also includes the effect of the conductivity (diffusion of information) between neighbouring sensors, which is modeled by an exponential decay factor, thus:

$$\hat{h}_S(\eta_1, \eta_2) = \operatorname{sinc}(\eta_1 c)\operatorname{sinc}(\eta_2 c) e^{-2\pi\beta_1 c|\eta_1|} e^{-2\pi\beta_2 c|\eta_2|},$$

where $\operatorname{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_1)$ and $\beta_1, \beta_2 > 0$.

**Optical system.** It is considered as an isotropic low-pass filter

$$\hat{h}_O(\eta_1, \eta_2) = e^{-2\pi\alpha c\sqrt{\eta_1^2+\eta_2^2}}, \quad \alpha > 0.$$

**Motion.** Each sensor counts the number of photons that arrive to its sensitive region during a certain time of acquisition. During the sampling time the system moves a distance $\tau$ and so does the sensor; this produces a motion blur effect in the motion direction $(d_1, d_2)$:

$$\hat{h}_M(\eta_1, \eta_2) = \operatorname{sinc}(\langle(\eta_1, \eta_2), (d_1, d_2)\rangle\tau).$$

Finally, the global MTF is the product of each of these intermediate transfer functions modeling the different aspects of the satellite:

$$\hat{h}(\eta_1, \eta_2) = \hat{h}_S\hat{h}_O\hat{h}_M.$$

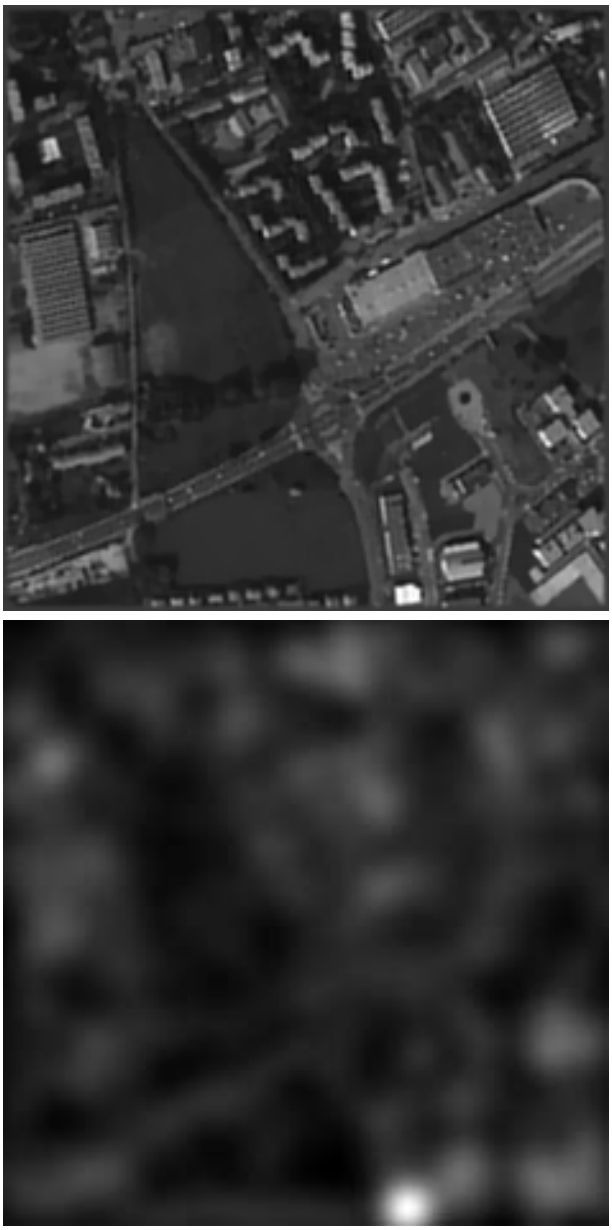


Figure 5. *Restored image with local Lagrange multipliers.* a) Top: the restored image corresponding to the data given in Figure 4.b. The restoration has been obtained using the algorithm of last section We have used a Gaussian function $G$ of standard deviation $\sigma = 6$. b) Bottom: the function $\lambda(i, j)$ obtained.

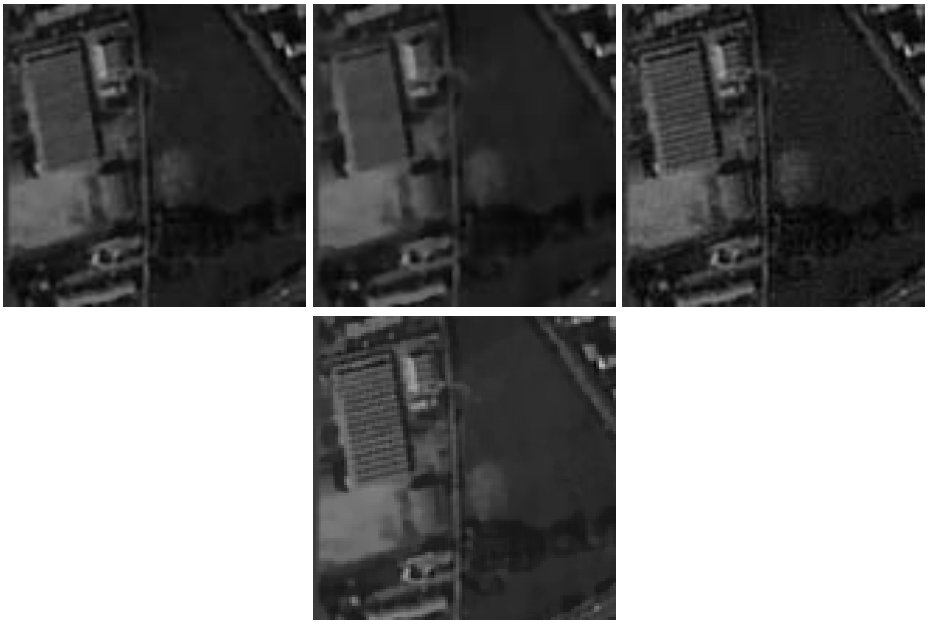


Figure 6. *A detail of the restored images with global and local constraints.* Top: a), b) and c) display a detail of the results that are obtained using a single global constraint (11) and show its main drawbacks. Figure a) corresponds to the result obtained with the value of $\lambda$ such that the constraint (11) is satisfied, in our case $\lambda = 10$. Figure b) shows that most textures are lost when using a small value of $\lambda$ ($\lambda = 2$)and Figure c) shows that some noise is present if we use a larger value of $\lambda$ ($\lambda = 1000$). Bottom: d) displays the same detail of Figure 5.a which has been obtained using restoration with local constraints.

**Acknowledgements.** The author is indebted to his coauthors: A. Almansa, F. Alter, F. Andreu, C. Ballester, G. Bellettini, M. Bertalmío, A. Chambolle, J. Faro, G. Haro, J. M. Mazón, M. Novaga, B. Rougé, and A. Solé. I would like to thank also J. F. Aujol and A. Chambolle for providing me with the images in Figure 3. Special thanks and a warm dedication to Andrés Solé who passed away after completing his PhD and started with me his work on restoration.

## References

[1] Almansa, A., Caselles, V., Haro, G., and Rougé, B., Restoration and zoom of irregularly sampled, blurred and noisy images by accurate total variation minimization with local constraints. *Multiscale Model. Simul.*, to appear.

[2] Almansa, A., C. Ballester, C., Caselles, V., Faro, L. J., and Haro, G., A TV based restoration model with local constraints. Preprint, 2006.

[3] Alter, F., Caselles, V., and Chambolle, A., Evolution of Characteristic Functions of Convex Sets in the Plane by the Minimizing Total Variation Flow. *Interfaces Free Bound.* **7** (2005), 29–53.

[4] Alter, F., Caselles, V., and Chambolle, A., A characterization of Convex Calibrable Sets in $\mathbb{R}^N$. *Math. Ann.* **332** (2005), 329–366.

[5] Ambrosio, L., Fusco, N., and Pallara, D., *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York 2000

[6] Andreu, F., Ballester, C., Caselles, V., and Mazón, J. M., Minimizing total variation flow. *Differential Integral Equations* **14** (2001), 321–360.

[7] Andreu-Vaillo, F., Caselles, V., and Mazón, J. M., *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals.* Progr. Math. 223, Birkhäuser, Basel 2004

[8] Andrews, H. C., and Hunt, B. R., *Digital Image Restoration.* Prentice Hall, Englewood Cliffs, NJ, 1977.

[9] Aujol, J. F., and Chambolle, A., Dual Norms and Image Decomposition Models. *Internat. J. Computer Vision*, **63** (2005), 85–104.

[10] Aujol, J. F., Aubert, G., Blanc-Feraud, L., and Chambolle, A., Image Decomposition into a bounded variation component and an oscillating component. *J. Math. Imaging Vision* **22** (2005), 71–88.

[11] Anzellotti, G., Pairings between measures and bounded functions and compensated compactness. *Ann. Mat. Pura Appl.* **135** (1983), 293–318.

[12] Bellettini, G., Caselles, V., and Novaga, M., The Total Variation Flow in $R^N$. *J. Differential Equations* **184** (2002), 475–525.

[13] Bellettini, G., Caselles, V., and Novaga, M., Explicit solutions of the eigenvalue problem $-\mathrm{div}\left(\frac{Du}{|Du|}\right) = u$. *SIAM J. Math. Anal.* **36** (2005), 1095–1129.

[14] Bermúdez, A., and Moreno, C., Duality methods for solving variational inequalities. *Comput. Math. Appl.* **7** (1981), 43–58.

[15] Bertalmío, M., Caselles, V., Rougé, B., and Solé, A., TV based image restoration with local constraints. *J. Sci. Comput.* **19** (2003), 95–122.

[16] Chambolle, A., An algorithm for total variation minimization and applications. *J. Math. Imaging Vision* **20** (2004), 89–97.

[17] Chambolle, A., and Lions, P. L., Image recovery via total variation minimization and related problems. *Numer. Math.* **76** (1997), 167–188.

[18] Chan, T. F., Golub, G. H., and Mulet, P., A Nonlinear Primal-Dual Method for Total Variation Based Image Restoration. *SIAM J. Sci. Comput.* **20** (1999), 1964–1977.

[19] Demoment, G., Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Trans. Acoust. Speech Signal Process.* **37** (1989), 2024–2036.

[20] Donoho, D., Denoising via soft-threholding. *IEEE Trans. Inform. Theory* **41** (1995), 613–627.

[21] Durand, S., Malgouyres, F., and Rougé, B., Image Deblurring, Spectrum Interpolation and Application to Satellite Imaging. *ESAIM Control Optim. Calc. Var.* **5** (2000), 445–475.

[22] Geman, D., and Reynolds, G., Constrained Image Restoration and Recovery of Discontinuities. *IEEE Trans. Pattern Anal. Machine Intell.* **14** (1992), 367–383.

[23] Gilboa, G., Sochen, N., and Zeevi, Y., PDE-based denoising of complex scenes using a spatially-varying fidelity term. In *Proc. International Conference on Image Processing 2003, Barcelona, Spain*, Vol. 1, 2003, 865–868.

[24] Giusti, E., On the equation of surfaces of prescribed mean curvature. Existence and uniqueness without boundary conditions. *Invent. Math.* **46** (1978), 111–137.

[25] Haddad, A., and Meyer, Y., Variational methods in image processing. *CAM Reports* 04-52, 2004.

[26] Meyer, Y., *Oscillating patterns in image processing and in some nonlinear evolution equations*. The Fifteenth Dean Jacqueline B. Lewis memorial lectures, University Lecture Series 22, Amer. Math. Soc., Providence, RI, 2001.

[27] Nikolova, M., Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math* **61** (2000), 633–658.

[28] Osher, S. J., Sole, A., and Vese, L. A., Image decomposition and restoration using total variation minimization and the $H^{-1}$ norm. *Multiscale Model. Simul.* **1** (2003), 349–370.

[29] Rougé, B., Théorie de l'echantillonage et satellites d'observation de la terre. In *Analyse de Fourier et traitement d'images*, Journées X-UPS 1998.

[30] Rudin, L., Osher, S., and Fatemi, E., Nonlinear total variation based noise removal algorithms. *Physica D* **60** (1992), 259–268.

[31] Tikhonov, A. N., and Arsenin, V. Y., *Solutions of Ill-Posed Problems*. Scripta Series in Mathematics, John Wiley & Sons, New York 1977.

[32] Vese, L. A., and Osher, S. J., Modeling textures with total variation mkinimization and oscillatikng patterns in image processing. *J. Sci. Comput.* **19** (2003), 553–572.

Departament Tecnologia, Universitat Pompeu Fabra, Passeig de Circumvalació, 8,
08003 Barcelona, Spain

E-mail: vicent.caselles@upf.edu

# A wavelet based sparse grid method for the electronic Schrödinger equation

Michael Griebel and Jan Hamaekers*

**Abstract.** We present a direct discretization of the electronic Schrödinger equation. It is based on one-dimensional Meyer wavelets from which we build an anisotropic multiresolution analysis for general particle spaces by a tensor product construction. We restrict these spaces to the case of antisymmetric functions. To obtain finite-dimensional subspaces we first discuss semi-discretization with respect to the scale parameter by means of sparse grids which relies on mixed regularity and decay properties of the electronic wave functions. We then propose different techniques for a discretization with respect to the position parameter. Furthermore we present the results of our numerical experiments using this new generalized sparse grid methods for Schrödinger's equation.

**Mathematics Subject Classification (2000).** 35J10, 65N25, 65N30, 65T40, 65Z05.

**Keywords.** Schrödinger equation, numerical approximation, sparse grid method, antisymmetric sparse grids.

## 1. Introduction

In this article we consider the electronic Schrödinger equation (first without spin for reasons of simplicity)

$$H\Psi(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) = E\Psi(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \tag{1}$$

with the Hamilton operator

$$H = T + V \quad \text{where } T = -\frac{1}{2}\sum_{p=1}^{N}\Delta_p$$

and

$$V = -\sum_{p=1}^{N}\sum_{q=1}^{N_{\text{nuc}}}\frac{Z_q}{|\boldsymbol{x}_p - \boldsymbol{R}_q|_2} + \sum_{p=1}^{N}\sum_{q>p}^{N}\frac{1}{|\boldsymbol{x}_p - \boldsymbol{x}_q|_2}. \tag{2}$$

*The authors were supported in part by the priority program 1145 *Modern and universal first-principles methods for many-electron systems in chemistry and physics* and the Sonderforschungsbereich 611 *Singuläre Phänomene und Skalierung in Mathematischen Modellen* of the Deutsche Forschungsgemeinschaft.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Here, with $d = 3$, $\boldsymbol{x}_p := (x_{1,p}, \ldots, x_{d,p}) \in \mathbb{R}^d$ denotes the position of the $p$-th electron, $p = 1 \ldots, N$, and $\boldsymbol{R}_q \in \mathbb{R}^d$ denotes the fixed position of the $q$-th nucleus, $q = 1, \ldots, N_{\text{nuc}}$. The operator $\Delta_p$ is the Laplacian acting on the $\boldsymbol{x}_p$-component of $\Psi$, i.e. $\Delta_p = \sum_{i=1}^{d} \partial^2/\partial(x_{i,p})^2$, $Z_q$ is the charge of the $q$-th nucleus and the norm $|\,.\,|_2$ denotes the usual Euclidean distance in $\mathbb{R}^d$. The solution $\Psi$ describes the wave function associated to the eigenvalue $E$.

This eigenvalue problem results from the Born–Oppenheimer approximation [51] to the general Schrödinger equation for a system of electrons and nuclei which takes the different masses of electrons and nuclei into account. It is one of the core problems of computational chemistry. Its successful treatment would allow to predict the properties of arbitrary atomic systems and molecules [22]. However, except for very simple cases, there is no analytical solution for (1) available. Also a direct numerical approach is impossible since $\Psi$ is a $d \cdot N$-dimensional function. Any discretization on e.g. uniform grids with $O(K)$ points in each direction would involve $O(K^{d \cdot N})$ degrees of freedoms which are impossible to store for $d = 3$, $N > 1$. Here, we encounter the curse of dimensionality [8]. Therefore, most approaches resort to an approximation of (1) only. Examples are the classical Hartree–Fock method and its successive refinements like configuration interaction and coupled clusters. Alternative methods are based on density functional theory which result in the Kohn–Sham equations or the reduced density matrix (RDM) [50] and the r12 approach [23] which lead to improved accuracy and open the way to new applications. A survey of these methods can be found in [3], [10], [46]. A major problem with these techniques is that, albeit quite successful in practice, they nevertheless only provide approximations. A systematical improvement is usually not available such that convergence of the model to Schrödinger's equation is achieved.

Instead, we intend to directly discretize the Schrödinger equation without resorting to any model approximation. To this end, we propose a new variant of the so-called sparse grid approach. The sparse grid method is a discretization technique for higher-dimensional problems which promises to circumvent the above-mentioned curse of dimensionality provided that certain smoothness prerequisites are fulfilled. Various sparse grid discretization methods have already been developed in the context of integration problems [27], [28], integral equations [24], [32] and elliptic partial differential equations, see [12] and the references cited therein for an overview. In Fourier space, such methods are also known under the name hyperbolic cross approximation [5], [21], [61]. A first heuristic approach to apply this methodology to the electronic Schrödinger equation was presented in [26]. The sparse grid idea was also used in the fast evaluation of Slater determinants in [33]. Recently Yserentant showed in [67] that the smoothness prerequisite necessary for sparse grids is indeed valid for the solution of the electronic Schrödinger equation. To be more precise, he showed that an antisymmetric solution of the electronic Schrödinger equation with $d = 3$ possesses $\mathcal{H}^{1,1}_{\text{mix}}$- or $\mathcal{H}^{1/2,1}_{\text{mix}}$-regularity for the fully antisymmetric and the partially symmetric case, respectively. This motivated the application of a generalized sparse grid approach in Fourier space to the electronic Schrödinger equation as presented

in [30]. There, sparse grids for general particle problems as well as antisymmetric sparse grids have been developed and were applied to (1) in the periodic setting. Basically, estimates of the type

$$\|\Psi - \Psi_M\|_{\mathcal{H}^1} \leq C(N,d) \cdot M^{-1/d} \cdot \|\Psi\|_{\mathcal{H}^{1,1}_{\text{mix}}}$$

could be achieved where $M$ denotes the number of Fourier modes used in the discretization. Here, the norm $\|.\|_{\mathcal{H}^{1,1}_{\text{mix}}}$ involves bounded mixed first derivatives. Thus the order of the method with respect to $M$ is asymptotically independent of the dimension of the problem, i.e. the number $N$ of electrons. But, the constants and the $\mathcal{H}^{1,1}_{\text{mix}}$-norm of the solution nevertheless depend on the number of electrons. While the dependency of the order constant might be analysed along the lines of [29], the problem remains that the smoothness term $\|\Psi\|_{\mathcal{H}^{1,1}_{\text{mix}}}$ grows exponentially with the number of electrons. This could be seen from the results of the numerical experiments in [30] and was one reason why in the periodic Fourier setting problems with higher numbers of electrons could not be treated. It was also observed in [69] where a certain scaling was introduced into the definitions of the norms which compensates for this growth factor. In [68], [70] it was suggested to scale the decomposition of the hyperbolic cross into subspaces accordingly and to further approximate each of the subspace contributions by some individually properly truncated Fourier series to cope with this problem.

In this article, we present a modified sparse grid/hyperbolic cross discretization for the electronic Schrödinger equation which implements this approach. It uses one-dimensional Meyer wavelets as basic building blocks in a tensor product construction to obtain a $\mathcal{L}^2$-orthogonal multiscale basis for the many-electron space. Then a truncation of the associated series expansion results in sparse grids. Here, for the level index we truncate according to the idea of hyperbolic crosses whereas we truncate for the position index according to various patterns which take to some extent the decay of the scaling function coefficients for $x \to \infty$ into account. Note that since we work in an infinite domain this resembles a truncation to a compact domain in which we then consider a local wavelet basis. Here, domain truncation error and scale resolution error should be balanced. Antisymmetry of the resulting discrete wavelet basis is achieved by a restriction of the active indices.

The remainder of this article is organized as follows: In Section 2 we present the Meyer wavelet family on $\mathbb{R}$ and discuss its properties. In Section 3 we introduce a multiresolution analysis for many particle spaces build by a tensor product construction from the one-dimensional Meyer wavelets and introduce various Sobolev norms. Then we discuss semi-discretization with respect to the scale parameter by means of generalized sparse grids and present a resulting error estimate in Section 4. Section 5 deals with antisymmetric generalized sparse grids. In Section 6 we invoke results on the mixed regularity of electronic wave functions and we discuss rescaling of norms and sparse grid spaces to obtain error bounds which involve the $\mathcal{L}^2$-norm of the solution instead of the mixed Sobolev norm. Then, in Section 7 we comment on the setup

of the system matrix and on the solution procedure for the discrete eigenvalue problem on general sparse grids and we propose different techniques for the discretization with respect to the position parameter. Furthermore we present the results of our numerical experiments. Finally we give some concluding remarks in Section 8.

## 2. Orthogonal multilevel bases and the Meyer wavelet family on $\mathbb{R}$

We intend to use for the discretization of (1) a $\mathcal{L}^2$-orthogonal basis system.[1] This is an important prerequisite from the practical point of view, since it allows to apply the well-known Slater–Condon rules. They reduce the $\mathbb{R}^{d\cdot N}$- and $\mathbb{R}^{2\cdot d\cdot N}$-dimensional integrals necessary in the Galerkin discretization of the one- and two electron part of the potential function of (1) to short sums of $d$-dimensional and $2d$-dimensional integrals, respectively. Otherwise, due to the structure of the Slater determinants necessary to obtain antisymmetry, these sums would contain exponentially many terms with respect to the number $N$ of electrons present in the system.

Let us recall the definition of a multiresolution analysis on $\mathbb{R}$, see also [52]. We consider an infinite sequence

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$$

of nested spaces $V_l$ with $\bigcap_{l\in\mathbb{Z}} V_l = 0$ and $\overline{\bigcup_{l\in\mathbb{Z}} V_l} = \mathcal{L}^2(\mathbb{R})$. It holds $f(x) \in V_l \Leftrightarrow f(2x) \in V_{l+1}$ and $f(x) \in V_0 \Leftrightarrow f(x-j) \in V_0$, where $j \in \mathbb{Z}$. Furthermore, there is a so-called scaling function (or father wavelet) $\phi \in V_0$, such that $\{\phi(x-j) \,:\, j \in \mathbb{Z}\}$ forms an orthonormal basis for $V_0$. Then

$$\{\phi_{l,j}(x) = 2^{\frac{l}{2}}\phi(2^l x - j) \,:\, j \in \mathbb{Z}\}$$

forms an orthonormal basis of $V_l$ and we can represent any $u(x) \in V_l$ as $u(x) = \sum_{j=-\infty}^{\infty} v_{l,j}\phi_{l,j}(x)$ with coefficients $v_{l,j} := \int_{\mathbb{R}} \phi^*_{l,j}(x)u(x)dx$. With the definition

$$W_l \perp V_l,\ V_l \oplus W_l = V_{l+1} \tag{3}$$

we obtain an associated sequence of detail spaces $W_l$ with associated mother wavelet $\varphi \in W_0$, such that $\{\varphi(x-j) \,:\, j \in \mathbb{Z}\}$ forms an orthonormal basis for $W_0$. Thus

$$\{\varphi_{l,j}(x) = 2^{\frac{l}{2}}\varphi(2^l x - j) \,:\, j \in \mathbb{Z}\}$$

gives an orthonormal basis for $W_l$ and $\{\varphi_{l,j} \,:\, l, j \in \mathbb{Z}\}$ is an orthonormal basis of $\mathcal{L}^2(\mathbb{R})$. Then, we can represent any $u(x)$ in $\mathcal{L}^2(\mathbb{R})$ as

$$u(x) = \sum_{l=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} u_{l,j}\varphi_{l,j}(x) \tag{4}$$

[1] Note that a bi-orthogonal system would also work here.

with the coefficients $u_{l,j} := \int_{\mathbb{R}} \varphi^*_{l,j}(x)u(x)dx$.

In the following we focus on the Meyer wavelet family for the choice of $\phi$ and $\varphi$. There, with the definition of the Fourier transform $\mathcal{F}[f](\omega) = \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx$ we set as father and mother wavelet in Fourier space

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2\pi}} \begin{cases} 1 & \text{for } |\omega| \leq \frac{2}{3}\pi, \\ \cos(\frac{\pi}{2}\nu(\frac{3}{2\pi}|\omega| - 1)) & \text{for } \frac{2\pi}{3} < |\omega| \leq \frac{4\pi}{3}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

$$\hat{\varphi}(\omega) = \frac{1}{\sqrt{2\pi}} e^{-i\frac{\omega}{2}} \begin{cases} \sin(\frac{\pi}{2}\nu(\frac{3}{2\pi}|\omega| - 1)) & \text{for } \frac{2}{3}\pi \leq |\omega| \leq \frac{4}{3}\pi, \\ \cos(\frac{\pi}{2}\nu(\frac{3}{4\pi}|\omega| - 1)) & \text{for } \frac{4\pi}{3} < |\omega| \leq \frac{8\pi}{3}, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $\nu\colon \mathbb{R} \to \mathbb{R} \in C^r$ is a parameter function still do be fixed, which has the properties $\nu(x) = 0$ for $x \leq 0$, $\nu(x) = 1$ for $x > 1$ and $\nu(x) + \nu(1 - x) = 1$. By dilation and translation we obtain

$$\mathcal{F}[\phi_{l,j}](\omega) = \hat{\phi}_{l,j}(\omega) = 2^{-\frac{l}{2}} e^{-i2^{-l}j\omega} \hat{\phi}(2^{-l}\omega),$$
$$\mathcal{F}[\varphi_{l,j}](\omega) = \hat{\varphi}_{l,j}(\omega) = 2^{-\frac{l}{2}} e^{-i2^{-l}j\omega} \hat{\varphi}(2^{-l}\omega)$$

where the $\hat{\phi}_{l,j}$ and $\hat{\varphi}_{l,j}$ denote the dilates and translates of (5) and (6), respectively.

This wavelet family can be derived from a partition of unity $\sum_l \hat{\chi}_l(\omega) = 1$ for all $\omega \in \mathbb{R}$ in Fourier space, where

$$\hat{\chi}_l(\omega) = \begin{cases} 2\pi \hat{\phi}^*_{0,0}(\omega)\hat{\phi}_{0,0}(\omega) & \text{for } l = 0, \\ 2^l \pi \hat{\varphi}^*_{l-1,0}(\omega)\hat{\varphi}_{l-1,0}(\omega) & \text{for } l > 0, \end{cases} \tag{7}$$

see [4] for details. The function $\nu$ basically describes the decay from one to zero of one partition function $\hat{\chi}_l$ in the overlap with its neighbor. The smoothness of the $\hat{\chi}_l$ is thus directly determined by the smoothness of $\nu$. The mother wavelets $\hat{\varphi}_{l,j}$ and the father wavelets $\hat{\phi}_{l,j}$ in Fourier space inherit the smoothness of the $\hat{\chi}_l$'s via the relation (7).

There are various choices for $\nu$ with different smoothness properties in the literature, see [4], [45], [53], [54]. Examples are the Shannon wavelet and the raised cosine wavelet [63], i.e. (6) with

$$\nu(x) = \nu^0(x) := \begin{cases} 0 & \text{for } x \leq \frac{1}{2}, \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad \nu(x) = \nu^1(x) \begin{cases} 0 & \text{for } x \leq 0, \\ x & \text{for } 0 \leq x \leq 1, \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

or, on the other hand,

$$\nu(x) = \nu^{\infty}(x) := \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{\tilde{\nu}(x)}{\tilde{\nu}(1-x)+\tilde{\nu}(x)} & \text{for } 0 < x \leq 1 \\ 1 & \text{otherwise} \end{cases} \quad \text{where } \tilde{\nu}(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ e^{-\frac{1}{x^{\alpha}}} & \text{otherwise} \end{cases} \tag{9}$$

with $\alpha = 1, 2$ [62], respectively. Other types of Meyer wavelets with different smoothness properties can be found in [19], [34], [40], [65]. The resulting mother wavelet functions in real space and Fourier space are given in Figure 1. Note the

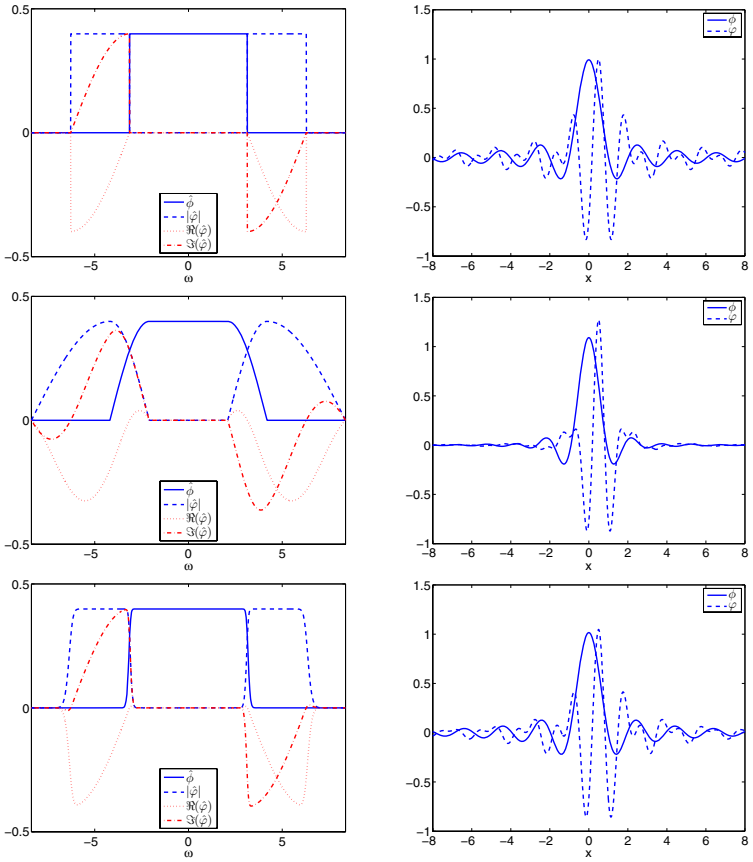



Figure 1. Top: (6) with $\nu^0$ from (8) in Fourier space (left) and real space (right). Middle: (6) with $\nu^1$ from (8) in Fourier space (left) and real space (right). Bottom: (6) with $\nu^{\infty}$ from (9) in Fourier space (left) and real space (right).

two symmetric areas of support and the associated two bands with non-zero values of the wavelets in Fourier space which resemble the line of construction due to Wilson,

Malvar, Coifman and Meyer [17], [20], [39], [49], [64] to circumvent the Balian–Low theorem[2] [7], [48]. In real space, these wavelets are $C^\infty$-functions with global support, in Fourier space, they are piecewise continuous, piecewise continuous differentiable and $C^\infty$, respectively, and have compact support. Furthermore they possess infinitely many vanishing moments. Finally their envelope in real space decays with $|x| \to \infty$ as $|x|^{-1}$ for $\nu^0$, as $|x|^{-2}$ for $\nu^1$ and faster than any polynomial (subexponentially) for $\nu^\infty$, respectively. To our knowledge, only for the Meyer wavelets with (8) there are analytical formulae in both real and Fourier space available. Certain integrals in a Galerkin discretization of (1) can then be given analytically. For the other types of Meyer wavelets analytical formulae only exist in Fourier space and thus numerical integration is necessary in a Galerkin discretization of (1).

For a discretization of (4) with respect to the level-scale $l$ we can restrict the doubly infinite sum to an interval, i.e. $l \in [L_1, L_2]$. However to obtain the space $V_{L_2}$ we have to complement the sum of detail spaces $W_l$, $l \in [L_1, L_2]$ by the space $V_{L_1}$, i.e. we have

$$V_{L_2} = V_{L_1} \oplus \bigoplus_{l=L_1}^{L_2} W_l.$$

with the associated representation

$$u(x) = \sum_{j=-\infty}^{\infty} v_{L_1,j}\phi_{L_1,j}(x) + \sum_{L_1}^{L_2}\sum_{j=-\infty}^{\infty} u_{l,j}\varphi_{l,j}(x).$$

Note that for the case of $\mathbb{R}$, beside the choice of a finest scale $L_2$, we here also have a choice of the coarsest scale $L_1$. This is in contrast to the case of a finite domain where the coarsest scale is usually determined by the size of the domain and is denoted as level zero.

Additionally we can scale our spaces and decompositions by a parameter $c > 0$, $c \in \mathbb{R}$. For example, we can set

$$V_l^c = \operatorname{span}\{\phi_{c,l,j}(x) = c^{\frac{1}{2}} 2^{\frac{l}{2}} \phi(c2^l x - j) \ : \ j \in \mathbb{Z}\}.$$

For $c = 2^k$, $k \in \mathbb{Z}$, the obvious identity $V_l^c = V_{l+k}^1$ holds. Then we obtain the scaled decomposition

$$V_{L_2}^c = V_{L_1}^c \oplus \bigoplus_{l=L_1}^{L_2} W_l^c$$

with the scaled detail spaces $W_l^c = \operatorname{span}\{\varphi_{c,l,j}(x) = c^{\frac{1}{2}} 2^{\frac{l}{2}} \varphi(c2^l x - j) \ : \ j \in \mathbb{Z}\}$. For $c = 2^k$, $k \in \mathbb{Z}$, the identity $W_l^c = W_{l+k}^1$ holds.

[2]The Balian–Low theorem basically states that the family of functions $g_{m,n}(x) = e^{2\pi i m x} g(x-n), m, n \in \mathbb{Z}$, which are related to the windowed Fourier transform, cannot be an orthonormal basis of $\mathcal{L}^2(\mathbb{R})$, if the two integrals $\int_{\mathbb{R}} x^2 |g(x)|^2 dx$ and $\int_{\mathbb{R}} k^2 |\hat{g}(k)|^2 dk$ are both finite. Thus there exists no orthonormal family for a Gaussian window function $g(x) = \pi^{-1/4} e^{-x^2/2}$ which is both sufficiently regular and well localized.

With the choice $c = 2^{-L_1}$ we can get rid of the parameter $L_1$ and may write our wavelet decomposition as

$$V_L^c = V_0^c \oplus \bigoplus_{l=0}^{L} W_l^c, \tag{10}$$

i.e. we can denote the associated coarsest space with level zero and the finest detail space with level $L$ (which now expresses the rescaled parameter $L_2$). To simplify notation we will skip the scaling index $c$ in the following.

We also introduce with

$$\psi_{l,j} := \begin{cases} \phi_{l,j}^c & \text{for } l = 0, \\ \varphi_{l-1,j}^c & \text{for } l \geq 1 \end{cases} \tag{11}$$

for $c = 2^{-L_1}$ a unique notation for both the father wavelets on the coarsest scale and the mother wavelets of the detail spaces. Bear however in mind that in the following the function $\psi_{l,j}$ with $l = 0$ denotes a father wavelet, i.e. a scaling function only, whereas it denotes for $l \geq 1$ a true wavelet on scale $l - 1$.

Let us finally consider the wavelet representation of the function $e^{-\sigma|x-x_0|}$ which is the one-dimensional analogon of the ground state wavefunction of hydrogen centered in $x_0 = 0$. For two types of Meyer wavelets, i.e. with $\nu^0$ from (8) and $\nu^\infty$ from (9) with $\alpha = 2$, Figure 2 gives the isolines to the values $10^{-3}$ and $10^{-4}$ for both the absolute value of the coefficients $v_{l,j}$ of the representation with respect to the scaling functions and the absolute value of the coefficients $u_{l,j}$ of the representation with respect to the wavelet functions.

Here we see the following: For the Meyer wavelet with $\nu^\infty$ from (9) where $\alpha = 2$, the isolines to different values (only $10^{-3}$ and $10^{-4}$ are shown) are nearly parallel for both the wavelet coefficients $u_{l,j}$ and the scaling coefficients $v_{l,j}$. For levels larger than $-2$ the isolines of the wavelet coefficients are even straight lines. Furthermore, on sufficiently coarse levels, the isoline for the wavelet coefficients and the scaling coefficients practically coincide. This is an effect of the $C^\infty$-property of the underlying mother wavelet. For the Meyer wavelet with $\nu^0$ from (8), i.e. for wavelets which are not $C^\infty$ in both real space *and* Fourier space, these two observations do not hold.

If we compare the isolines of the wavelet coefficients $u_{l,j}$ for the Meyer wavelet with $\nu^\infty$ from (9) where $\alpha = 2$ and that of the Meyer wavelet with $\nu^0$ from (8) we observe that the level on which the bottom kink occurs is exactly the same. However the size of the largest diameter (here roughly on level $-2$) is substantially bigger for the Shannon wavelet. Note the different scaling of the x-axis of the diagrams on the left and right side.

We furthermore observe for the isolines of the scaling coefficients an exponential behavior, i.e. from level $l$ to level $l + 1$ the associated value for $j$ nearly doubles in a sufficient distance away from point $x = 0$. With respect to the wavelet coefficients, however, we see that the support shrinks super-exponentially towards the bottom kink with raising level.

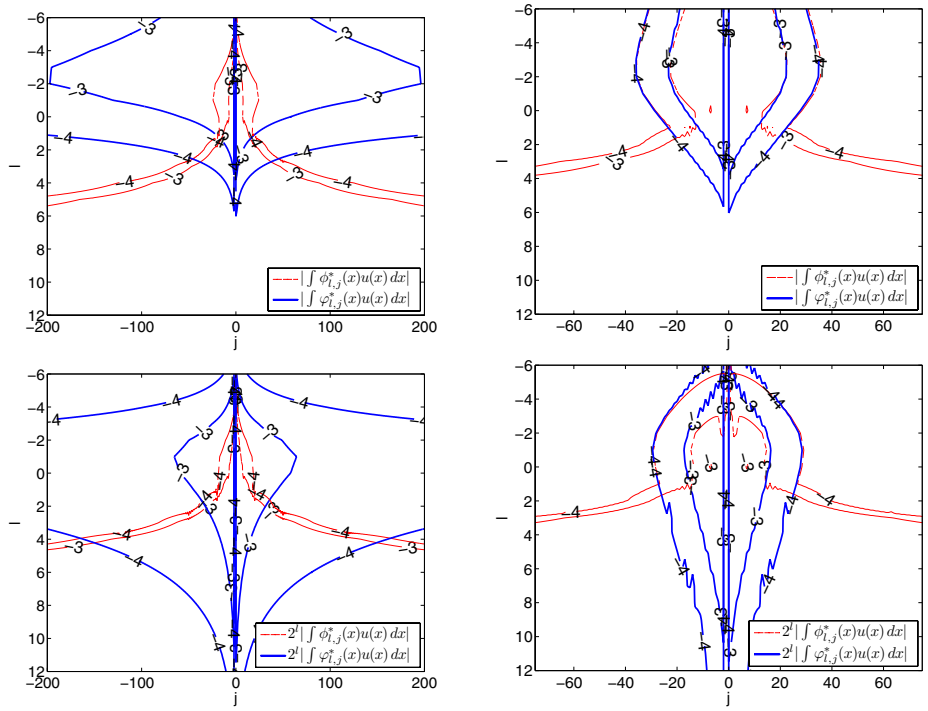



Figure 2. Isolines to the values $10^{-3}$ and $10^{-4}$ of the absolute value of the coefficients $v_{l,j}$ and $u_{l,j}$ for the Meyer wavelets with $\nu^0$ from (8) (left) and $\nu^\infty$ from (9) with $\alpha = 2$ (right), no scaling (top) and scaling with $2^l$ (bottom).

The relation (3) relates the spaces $V_l$, $W_l$ and $V_{l+1}$ and allows to switch between the scaling coefficients and the wavelet coefficients on level $l$ to the scaling coefficients on level $l+1$ and vice versa. This enables us to choose an optimal coarsest level for a prescribed accuracy and we also can read off the pattern of indices $(l, j)$ which result in a best $M$-term approximation with respect to the $\mathcal{L}^2$- and $\mathcal{H}^1$-norm for that prescribed accuracy, respectively. For the Meyer wavelet with $\nu^\infty$ from (9) where $\alpha = 2$, the optimal choice of the coarsest level $L_1$ on which we use scaling functions is just the level where, for a prescribed accuracy, the two absolute values of the wavelet coefficients on one level possess their largest distance, i.e. the associated isoline of the wavelet coefficients shows the largest diameter (here roughly on level $-2$). The selection of a crossing isoline then corresponds to the fixation of a boundary error by truncation of the further decaying scaling function coefficients on that level which resembles a restriction of $\mathbb{R}$ to just a finite domain. From this base a downward pointing triangle then gives the area of indices to be taken into account into the finite sum of best approximation with respect to that error. We observe that the use of the wavelets with $\nu^0$ from (8) would result in a substantially larger area of indices and

thus number of coefficients to be taken into account to obtain the same error level. There, the form of the area is no longer a simple triangle but shows a "butterfly"-like shape where the base of the pattern is substantially larger.

## 3. MRA and Sobolev spaces for particle spaces

In the following we introduce a multiresolution analysis based on Meyer wavelets for particle spaces on $(\mathbb{R}^d)^N$ and discuss various Sobolev spaces on it.

First, let us set up a basis for the one-particle space $\mathcal{H}^s(\mathbb{R}^d) \subset \mathcal{L}^2(\mathbb{R}^d)$. Here, we use the $d$-dimensional product of the one-dimensional system $\{\psi_{l,j}(x), l \in \mathbb{N}_0, j \in \mathbb{Z}\}$. We then define the $d$-dimensional multi-indices $\boldsymbol{l} = (l_1, l_2, \ldots, l_d) \in \mathbb{N}_0^d$ and $\boldsymbol{j} = (j_1, j_2, \ldots, j_d) \in \mathbb{Z}^d$, the coordinate vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and the associated $d$-dimensional basis functions

$$\psi_{\boldsymbol{l},\boldsymbol{j}}(\boldsymbol{x}) := \prod_{i=1}^{d} \psi_{l_i,j_i}(x_i). \tag{12}$$

Note that due to (11) this product may involve both father and mother wavelets depending on the values of the components of the level index $\boldsymbol{l}$. We furthermore denote $|\boldsymbol{l}|_2 = \left(\sum_{i=1}^d l_i^2\right)^{1/2}$ and $|\boldsymbol{l}|_\infty = \max_{1\le i\le d} |l_i|$. Let us now define isotropic Sobolev spaces in $d$ dimensions with help of the wavelet series expansion, i.e. we classify functions via the decay of their wavelet coefficients. To this end, we set

$$\lambda(\boldsymbol{l}) := |2^{\boldsymbol{l}}|_2 = |(2^{l_1}, \ldots, 2^{l_d})|_2 \tag{13}$$

and define

$$\begin{aligned}\mathcal{H}^s(\mathbb{R}^d) = \Big\{u(\boldsymbol{x}) = \sum\nolimits_{\substack{l\in\mathbb{N}_0^d,\\ \boldsymbol{j}\in\mathbb{Z}^d}} u_{\boldsymbol{l},\boldsymbol{j}}\psi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{x}) :\\ \|u\|^2_{\mathcal{H}^s(\mathbb{R}^d)} = \sum\nolimits_{\substack{l\in\mathbb{N}_0^d,\\ \boldsymbol{j}\in\mathbb{Z}^d}} \lambda(\boldsymbol{l})^{2s}\cdot|u_{\boldsymbol{l},\boldsymbol{j}}|^2 \le c^2 < \infty\Big\},\end{aligned} \tag{14}$$

where $u_{\boldsymbol{l},\boldsymbol{j}} = \int_{\mathbb{R}^d} \psi^*_{\boldsymbol{l},\boldsymbol{j}}(\boldsymbol{x})u(\boldsymbol{x})d\vec{\boldsymbol{x}}$ and $c$ is a constant which depends on $d$.

Based on the given one-particle basis (12) we now define a basis for many-particle spaces on $\mathbb{R}^{d\cdot N}$. We then have the $d \cdot N$-dimensional coordinates $\vec{\boldsymbol{x}} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ where $\boldsymbol{x}_i \in \mathbb{R}^d$. To this end, we first employ a tensor product construction and define the multi-indices $\vec{\boldsymbol{l}} = (\boldsymbol{l}_1, ..., \boldsymbol{l}_N) \in \mathbb{N}_0^{d\cdot N}$ and the associated multivariate wavelets

$$\psi_{\vec{\boldsymbol{l}},\vec{\boldsymbol{j}}}(\vec{\boldsymbol{x}}) := \prod_{p=1}^{N} \psi_{\boldsymbol{l}_p,\boldsymbol{j}_p}(\boldsymbol{x}_p) = \Big(\bigotimes_{p=1}^{N} \psi_{\boldsymbol{l}_p,\boldsymbol{j}_p}\Big)(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N). \tag{15}$$

Note again that this product may involve both father and mother wavelets depending on the values of the components of the level index $\vec{\boldsymbol{l}}$. The wavelets $\psi_{\vec{\boldsymbol{l}},\vec{\boldsymbol{j}}}$ span the

subspaces $W_{\vec{l},\vec{j}} := \operatorname{span}\{\psi_{\vec{l},\vec{j}}\}$ whose union forms[3] the space

$$V = \bigoplus_{\substack{\vec{l}\in\mathbb{N}_0^{dN} \\ \vec{j}\in\mathbb{Z}^{dN}}} W_{\vec{l},\vec{j}}. \tag{16}$$

We then can uniquely represent any function $u$ from $V$ as

$$u(\vec{x}) = \sum_{\substack{\vec{l}\in\mathbb{N}_0^{dN} \\ \vec{j}\in\mathbb{Z}^{dN}}} u_{\vec{l},\vec{j}}\ \psi_{\vec{l},\vec{j}}(\vec{x}) \tag{17}$$

with coefficients $u_{\vec{l},\vec{j}} = \int_{\mathbb{R}^{dN}} \psi^*_{\vec{l},\vec{j}}(\vec{x})u(\vec{x})d\vec{x}$.

Now, starting from the one-particle space $\mathcal{H}^s(\mathbb{R}^d)$ we build Sobolev spaces for many particles. Obviously there are many possibilities to generalize the concept of Sobolev spaces [1] from the one-particle case to higher dimensions. Two simple possibilities are the additive or multiplicative combination i.e. an arithmetic or geometric averaging of the scales for the different particles. We use the following definition that combines both possibilities. We denote

$$\lambda_{\text{mix}}(\vec{l}) := \prod_{p=1}^{N} \lambda(\boldsymbol{l}_p) \quad \text{and} \quad \lambda_{\text{iso}}(\vec{l}) := \sum_{p=1}^{N} \lambda(\boldsymbol{l}_p). \tag{18}$$

Now, for $-\infty < t, r < \infty$, set

$$\begin{aligned} &\mathcal{H}^{t,r}_{\text{mix}}((\mathbb{R}^d)^N) \\ &\quad = \Big\{u(\vec{x}) = \textstyle\sum_{\substack{\vec{l}\in\mathbb{N}_0^{dN} \\ \vec{j}\in\mathbb{Z}^{dN}}} u_{\vec{l},\vec{j}}\psi_{\vec{l},\vec{j}}(\vec{x}) : \\ &\qquad \|u\|^2_{\mathcal{H}^{t,r}_{\text{mix}}((\mathbb{R}^d)^N)} = \textstyle\sum_{\vec{l}\in\mathbb{N}_0^{dN}} \lambda_{\text{mix}}(\vec{l})^{2t} \cdot \lambda_{\text{iso}}(\vec{l})^{2r} \cdot \sum_{\vec{j}\in\mathbb{Z}^{dN}} |u_{\vec{l},\vec{j}}|^2 \le c^2 < \infty\Big\} \end{aligned} \tag{19}$$

with a constant $c$ which depends on $d$ and $N$.

The standard isotropic Sobolev spaces [1] as well as the Sobolev spaces of dominating mixed smoothness [58], both generalized to the $N$-particle case, are included here. They can be written as the special cases

$$\mathcal{H}^s((\mathbb{R}^d)^N) = \mathcal{H}^{0,s}_{\text{mix}}((\mathbb{R}^d)^N) \quad \text{and} \quad \mathcal{H}^t_{\text{mix}}((\mathbb{R}^d)^N) = \mathcal{H}^{t,0}_{\text{mix}}((\mathbb{R}^d)^N),$$

respectively. Hence, the parameter $r$ from (19) governs the isotropic smoothness, whereas $t$ governs the mixed smoothness. Thus, the spaces $\mathcal{H}^{t,r}_{\text{mix}}$ give us a quite flexible framework for the study of problems in Sobolev spaces. Note that the relations $\mathcal{H}^t_{\text{mix}} \subset \mathcal{H}^t \subset \mathcal{H}^{t/N}_{\text{mix}}$ for $t \ge 0$ and $\mathcal{H}^{t/N}_{\text{mix}} \subset \mathcal{H}^t \subset \mathcal{H}^t_{\text{mix}}$ for $t \le 0$ hold. See [58] and [36] for more information on the spaces $\mathcal{H}^t_{\text{mix}}$.

[3] Except for the completion with respect to a chosen Sobolev norm, $V$ is just the associated Sobolev space.

## 4. Semidiscrete general sparse grid spaces

We now consider truncation of the series expansion (17) with respect to the level parameter $\vec{l}$ but keep the part of the full series expansion with respect to the position parameter $\vec{j}$. To this end, we introduce, besides the parameter $L$ (after proper scaling with $c$) which indicates the truncation of the scale with respect to the one-particle space, an additional parameter $T \in (-\infty, 1]$ which regulates the truncation pattern for the interaction between particles. We define the generalized sparse grid space

$$V_{L,T} := \bigoplus_{\vec{l} \in \Omega_{L,T}} W_{\vec{l}} \quad \text{where } W_{\vec{l}} = \text{span}\{\psi_{\vec{l},\vec{j}}, \vec{j} \in \mathbb{Z}^{dN}\} \tag{20}$$

with associated generalized hyperbolic cross with respect to the scale-parameter $\vec{l}$

$$\Omega_{L,T} := \{\vec{l} \in \mathbb{N}_0^{d \cdot N} : \lambda_{\text{mix}}(\vec{l}) \cdot \lambda_{\text{iso}}(\vec{l})^{-T} \leq (2^L)^{1-T}\}. \tag{21}$$

The parameter $T$ allows us to switch from the full grid case $T = -\infty$ to the conventional sparse grid case $T = 0$, compare [12], [31], [42], and also allows to create with $T \in (0, 1]$ subspaces of the hyperbolic cross/conventional sparse grid space. Obviously, the inclusions $V_{L,T_1} \subset V_{L,T_2}$ for $T_1 \leq T_2$ hold. Figure 3 displays the index sets for various choices of $T$ for the case $d = 1$, $N = 2$ and $L = 128$.

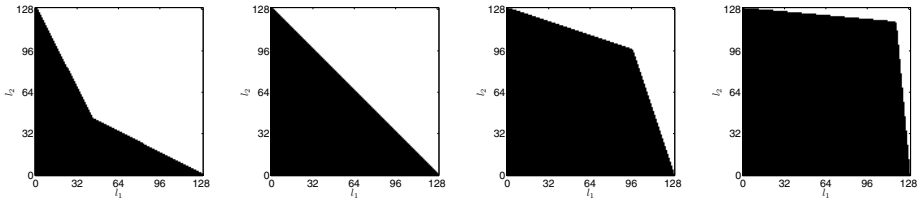



Figure 3. $\Omega_{128,T}$ for $T = 0.5, 0, -2, -10$ (from left to right), $d = 1$, $N = 2$; the conventional sparse grid/hyperbolic cross corresponds to $T = 0$. For $T = -\infty$ we get a completely black square.

We then can uniquely represent any function $u$ from $V_{K,T}$ as

$$u(\vec{x}) = \sum_{\vec{l} \in \Omega_{L,T}, \vec{j} \in \mathbb{Z}^{d \cdot N}} u_{\vec{l},\vec{j}} \; \psi_{\vec{l},\vec{j}}(\vec{x}).$$

Such a projection into $V_{K,T}$ introduces an error. Here we have the following error estimate:

**Lemma 1.** *Let $s < r + t, t \geq 0$, $u \in \mathcal{H}^{t,r}_{\text{mix}}((\mathbb{R}^d)^N)$. Let $\tilde{u}_{L,T}$ be the best approximation in $V_{L,T}$ with respect to the $\mathcal{H}^s$-norm and let $u_{L,T}$ be the interpolant of $u$ in*

$V_{L,T}$, *i.e.* $u_{L,T} = \sum_{\vec{l} \in \Omega_{L,T}} \sum_{\vec{j} \in \mathbb{Z}^{dN}} u_{\vec{l},\vec{j}} \psi_{\vec{l},\vec{j}}(\vec{x})$. *Then, there holds*

$$\inf_{V_{L,T}} \|u - v\|_{\mathcal{H}^s} = \|u - \tilde{u}_{L,T}\|_{\mathcal{H}^s} \leq \|u - u_{L,T}\|_{\mathcal{H}^s} \leq \begin{cases} O((2^L)^{s-r-t+(Tt-s+r)\frac{N-1}{N-T}}) \cdot \|u\|_{\mathcal{H}^{t,r}_{\mathrm{mix}}} & \text{for } T \geq \frac{s-r}{t}, \\ O((2^L)^{s-r-t}) \cdot \|u\|_{\mathcal{H}^{t,r}_{\mathrm{mix}}} & \text{for } T \leq \frac{s-r}{t}. \end{cases} \quad (22)$$

For a proof, compare the arguments in [31], [42], [43], [30]. This type of estimate was already given for the case of a dyadically refined wavelet basis with $d = 1$ for the periodic case on a finite domain in [31], [42], [43]. It is a generalization of the energy-norm based sparse grid approach of [11], [12], [29] where the case $s = 1$, $t = 2$, $r = 0$ was considered using a hierarchical piecewise linear basis.

Let us discuss some cases. For the standard Sobolev space $\mathcal{H}^{0,r}_{\mathrm{mix}}$ (i.e. $t = 0$, $r = 2$) and the spaces $V_{L,T}$ with $T \geq -\infty$ the resulting order is dependent of $T$ and dependent on the number of particles $N$. In particular the order even deteriorates with larger $T$. For the standard Sobolev spaces of bounded mixed derivatives $\mathcal{H}^{t,0}_{\mathrm{mix}}$ (i.e. $t = 2$, $r = 0$) and the spaces $V_{L,T}$ with $T > \frac{s}{2}$ the resulting order is dependent of $T$ and dependent on the number of particles $N$ whereas for $T \leq \frac{s}{2}$ the resulting order is independent of $T$ and $N$. If we restrict the class of functions for example to $\mathcal{H}^{1,1}_{\mathrm{mix}}$ (i.e. $t = 1$, $r = 1$) and measure the error in the $\mathcal{H}^1$-norm (i.e. $s = 1$) the approximation order is dependent on $N$ for all $T > 0$ and independent on $N$ and $T$ for all $T \leq 0$. Note that in all cases the constants in the $O$-notation depend on $N$ and $d$.

## 5. Antisymmetric semidiscrete general sparse grid spaces

Let us now come back to the Schrödinger equation (1). Note that in general an electronic wave function depends in addition to the positions $\boldsymbol{x}_i$ of the electrons also on their associated spin coordinates $\sigma_i \in \{-\frac{1}{2}, \frac{1}{2}\}$. Thus electronic wave functions are defined as $\Psi\colon (\mathbb{R}^d)^N \times \{-\frac{1}{2}, \frac{1}{2}\}^N \to \mathbb{R}\colon (\vec{x}, \vec{\sigma}) \to \Psi(\vec{x}, \vec{\sigma})$ with spin coordinates $\vec{\sigma} = (\sigma_1, \ldots, \sigma_N)$. Furthermore, physically relevant eigenfunctions $\Psi$ obey the following two assumptions: First, elementary particles are indistinguishable from each other (fundamental principle of quantum mechanics). Second, no two electrons may occupy the same quantum state simultaneously (Pauli exclusion principle). Thus, we consider only wave functions which are antisymmetric with respect to an arbitrary simultaneous permutation $P \in \mathcal{S}_N$, of the electron positions and spin variables, i.e. which fulfil

$$\Psi(P\vec{x}, P\vec{\sigma}) = (-1)^{|P|} \Psi(P\vec{x}, P\vec{\sigma}).$$

Here $\mathcal{S}_N$ is the symmetric group. The permutation $P$ is a mapping $P\colon \{1, \ldots, N\} \to \{1, \ldots, N\}$ which translates to a permutation of the corresponding numbering of

electrons and thus to a permutation of indices, i.e. we have $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T := (\boldsymbol{x}_{P(1)}, \ldots, \boldsymbol{x}_{P(N)})^T$ and $P(\sigma_1, \ldots, \sigma_N)^T := (\sigma_{P(1)}, \ldots, \sigma_{P(N)})^T$. In particular, the symmetric group is of size $|\mathcal{S}_N| = N!$ and the expression $(-1)^{|P|}$ is equal to the determinant $\det P$ of the associated permutation matrix.

Now, to a given spin vector $\vec{\sigma} \in \left\{-\frac{1}{2}, \frac{1}{2}\right\}^N$ we define the associated spatial component of the full wave function $\Psi$ by $\Psi_{\vec{\sigma}}\colon (\mathbb{R}^d)^N \to \mathbb{R}\colon \vec{\boldsymbol{x}} \to \Psi(\vec{\boldsymbol{x}}, \vec{\sigma})$. Then, since there are $2^N$ possible different spin distributions $\vec{\sigma}$, the full Schrödinger equation, i.e. the eigenvalue problem $H\Psi = E\Psi$, decouples into $2^N$ eigenvalue problems for the $2^N$ associated spatial components $\Psi_{\vec{\sigma}}$. Here, the spatial part $\Psi_{\vec{\sigma}}$ to a given $\vec{\sigma}$ obeys the condition

$$\Psi_{\vec{\sigma}}(P\vec{\boldsymbol{x}}) = (-1)^{|P|}\Psi_{\vec{\sigma}}(P\vec{\boldsymbol{x}}) \quad \text{for all } P \in \mathcal{S}_{\vec{\sigma}} := \{P \in \mathcal{S}_N \,:\, P\vec{\sigma} = \vec{\sigma}\}. \tag{23}$$

In particular, the minimal eigenvalue of all eigenvalue problems for the spatial components is equal to the minimal eigenvalue of the full eigenvalue problem. Moreover, the eigenfunctions of the full system can be composed by the eigenfunctions of the eigenvalue problems for the spatial parts.

Although there are $2^N$ possible different spin distributions $\vec{\sigma}$, the bilinear form $\langle \Psi(P\cdot)|H|\Psi(P\cdot)\rangle$ is invariant under all permutations $P \in \mathcal{S}_N$ of the position coordinates $\vec{\boldsymbol{x}}$. Thus it is sufficient to consider the eigenvalue problems which are associated to the spin vectors $\vec{\sigma}^{(N,S)} = (\sigma_1^{(N,S)}, \ldots, \sigma_N^{(N,S)})$ where the first $S$ electrons possess spin $-\frac{1}{2}$ and the remaining $N - S$ electrons possess spin $\frac{1}{2}$, i.e.

$$\sigma_j^{(N,S)} = \begin{cases} -\frac{1}{2} & \text{for } j \le S, \\ \frac{1}{2} & \text{for } j > S. \end{cases}$$

In particular, it is enough to solve only the $\lfloor N/2 \rfloor$ eigenvalue problems which correspond to the spin vectors $\vec{\sigma}^{(N,S)}$ with $S \le N/2$. For further details see [66]. Therefore, we consider in the following without loss of generality only spin distributions $\vec{\sigma}^{(N,S)} = (\sigma_1^{(N,S)}, \ldots, \sigma_N^{(N,S)})$. We set $\mathcal{S}_{(N,S)} := \mathcal{S}_{\vec{\sigma}^{(N,S)}}$. Note that there holds $|\mathcal{S}_{(N,S)}| = S!(N-S)!$.

Now we define spaces of antisymmetric functions and their semi-discrete sparse grid counterparts. The functions of the $N$-particle space $V$ from (16) which obey the anti-symmetry condition (23) for a given $\vec{\sigma}^{(N,S)}$ form a linear subspace $V^{\mathcal{A}^{(N,S)}}$ of $V$. We define the projection into this subspace, i.e. the antisymmetrization operator $\mathcal{A}^{(N,S)}\colon V \to V^{\mathcal{A}^{(N,S)}}$ by

$$\mathcal{A}^{(N,S)}u(\vec{\boldsymbol{x}}) := \frac{1}{S!(N-S)!} \sum_{P \in \mathcal{S}_{N,S}} (-1)^{|P|} u(P\vec{\boldsymbol{x}}). \tag{24}$$

For any basis function $\psi_{\vec{l},\vec{j}}$ of our general $N$-particle space $V$ we then have

$$\begin{aligned}
\mathcal{A}^{(N,S)}\psi_{\vec{l},\vec{j}}(\vec{x}) &= \mathcal{A}^{(N,S)}\Big(\Big(\bigotimes_{p=1}^{S}\psi_{l_p,j_p}\Big)(x_1,\ldots,x_S)\Big(\bigotimes_{p=S+1}^{N}\psi_{l_p,j_p}\Big)(x_{S+1},\ldots x_N)\Big)\\
&= \Big(\mathcal{A}^{(S,S)}\bigotimes_{p=1}^{S}\psi_{l_p,j_p}(x_1,\ldots,x_S)\Big)\Big(\mathcal{A}^{(N-S,N-S)}\bigotimes_{p=S+1}^{N}\psi_{l_p,j_p}(\vec{x}_{S+1},\ldots,x_N)\Big)\\
&= \Big(\frac{1}{S!}\bigwedge_{p=1}^{S}\psi_{l_p,j_p}\Big(x_1,\ldots,x_S\Big)\Big)\Big(\frac{1}{(N-S)!}\bigwedge_{p=S+1}^{N}\psi_{l_p,j_p}\Big(x_{S+1},\ldots,x_N\Big)\Big)\\
&= \frac{1}{S!(N-S)!}\sum_{P\in\mathcal{S}_{N,S}}(-1)^{|P|}\psi_{\vec{l},\vec{j}}(P\vec{x}) = \frac{1}{S!(N-S)!}\sum_{P\in\mathcal{S}_{N,S}}(-1)^{|P|}\psi_{P\vec{l},P\vec{j}}(\vec{x}).
\end{aligned}$$

In other words, the classical product

$$\psi_{\vec{l},\vec{j}}(\vec{x}) := \prod_{p=1}^{N}\psi_{l_p,j_p}(x_p) = \Big(\bigotimes_{p=1}^{N}\psi_{l_p,j_p}\Big)(x_1,\ldots,x_N)$$

gets replaced by the product of two outer products

$$\frac{1}{S!}\bigwedge_{p=1}^{S}\psi_{l_p,j_p}(x_1,\ldots,x_S) \quad\text{and}\quad \frac{1}{(N-S)!}\bigwedge_{p=S+1}^{N}\psi_{l_p,j_p}(x_{S+1},\ldots,x_N)$$

that correspond to the two sets of coordinates and one-particle bases which are associated to the two spin values $-\frac{1}{2}$ and $\frac{1}{2}$. The outer product involves just the so-called slater determinant [55], i.e.

$$\bigwedge_{p=1}^{N}\psi_{l_p,j_p}(x_1,\ldots,x_N) = \begin{vmatrix}\psi_{l_1,j_1}(x_1) & \ldots & \psi_{l_1,j_1}(x_N)\\ \vdots & \ddots & \vdots\\ \psi_{l_N,j_N}(x_1) & \ldots & \psi_{l_N,j_N}(x_N)\end{vmatrix}.$$

Note here again that due to (11) both father wavelet functions and mother wavelet functions may be involved in the respective products.

The sequence $\{\mathcal{A}^{(N,S)}\psi_{\vec{l},\vec{j}}\}_{\vec{l}\in\mathbb{N}_0^{dN},\vec{j}\in\mathbb{Z}^{dN}}$ only forms a generating system of the antisymmetric subspace $V^{\mathcal{A}^{(N,S)}}$ and no basis since many functions $\mathcal{A}^{(N,S)}\psi_{\vec{l},\vec{j}}$ are identical (up to the sign). But we can gain a basis for the antisymmetric subspace $V^{\mathcal{A}^{(N,S)}}$ if we restrict the sequence $\{\mathcal{A}^{(N,S)}\psi_{\vec{l},\vec{j}}\}_{\vec{l}\in\mathbb{N}_0^{dN},\vec{j}\in\mathbb{Z}^{dN}}$ properly. This can be done in many different ways. A possible orthonormal basis $\mathcal{B}^{(N,S)}$ for $V^{\mathcal{A}^{(N,S)}}$ is given with help of

$$\Phi^{(N,S)}_{\vec{l},\vec{j}}(\vec{x}) := \frac{1}{\sqrt{S!(N-S)!}}\cdot\bigwedge_{p=1}^{S}\psi_{l_p,j_p}(x_1,\ldots,x_S)\cdot\bigwedge_{p=S+1}^{N}\psi_{l_p,j_p}(x_{S+1},\ldots,x_N) \tag{25}$$

as follows:

$$\mathcal{B}^{(N,S)} := \big\{\Phi^{(N,S)}_{\vec{l},\vec{j}} : \vec{l} \in \mathbb{N}_0^{d\cdot N},\ \vec{j} \in \mathbb{Z}^{d\cdot N},\ (\boldsymbol{l}_1, \boldsymbol{j}_1) < \cdots < (\boldsymbol{l}_S, \boldsymbol{j}_S) \text{ and } (\boldsymbol{l}_{S+1}, \boldsymbol{j}_{S+1}) < \cdots < (\boldsymbol{l}_N, \boldsymbol{j}_N)\big\} \tag{26}$$

where for the index pair

$$\boldsymbol{I}_p := (\boldsymbol{l}_p, \boldsymbol{j}_p) = (\boldsymbol{l}_{p,(1)}, \ldots, \boldsymbol{l}_{p,(d)}, \boldsymbol{j}_{p,(1)}, \ldots, \boldsymbol{j}_{p,(d)})$$

the relation $<$ is defined as

$$\boldsymbol{I}_p < \boldsymbol{I}_q \quad \Longleftrightarrow \quad \begin{array}{l}\text{there exists } \alpha \in \{1, \ldots, 2d\} \text{ such that } \boldsymbol{I}_{p,(\alpha)} < \boldsymbol{I}_{q,(\alpha)} \\ \text{and } \boldsymbol{I}_{p,(\beta)} \le \boldsymbol{I}_{q,(\beta)} \text{ for all } \beta \in \{1, \ldots, \alpha - 1\}.\end{array}$$

With

$$\Omega^{\mathcal{A}^{(N,S)}} = \{(\vec{l}, \vec{j}) : \vec{l} \in \mathbb{N}_0^{d\cdot N},\ \vec{j} \in \mathbb{Z}^{d\cdot N},\ (\boldsymbol{l}_1, \boldsymbol{j}_1) < \cdots < (\boldsymbol{l}_S, \boldsymbol{j}_S) \text{ and } (\boldsymbol{l}_{S+1}, \boldsymbol{j}_{S+1}) < \cdots < (\boldsymbol{l}_N, \boldsymbol{j}_N)\big\}$$

we then can define the antisymmetric subspace $V^{\mathcal{A}^{(N,S)}}$ of $V$ as

$$V^{\mathcal{A}^{(N,S)}} = \bigoplus_{(\vec{l},\vec{j})\in\Omega^{\mathcal{A}^{(N,S)}}} W_{\vec{l},\vec{j}} \tag{27}$$

where we denote from now on $W_{\vec{l},\vec{j}} = \operatorname{span}\{\Phi^{(N,S)}_{\vec{l},\vec{j}}(\vec{x})\}$. Any function $u$ from $V^{\mathcal{A}^{(N,S)}}$ can then uniquely be represented as

$$u(\vec{x}) = \sum_{(\vec{l},\vec{j})\in\Omega^{\mathcal{A}^{(N,S)}}} u_{\vec{l},\vec{j}}\ \Phi^{(N,S)}_{\vec{l},\vec{j}}(\vec{x})$$

with coefficients $u_{\vec{l},\vec{j}} = \int_{I^{dN}} \Phi^{(N,S)*}_{\vec{l}\vec{j}}(\vec{x}) u(\vec{x}) d\vec{x}$.

Now we are in the position to consider semidiscrete subspaces of $V^{\mathcal{A}^{(N,S)}}$. To this end, in analogy to (20) we define the generalized semidiscrete antisymmetric sparse grid spaces

$$V^{\mathcal{A}^{(N,S)}}_{L,T} := \bigoplus_{(\vec{l},\vec{j})\in\Omega^{\mathcal{A}^{(N,S)}}_{K,T}} W_{\vec{l},\vec{j}}$$

with associated antisymmetric generalized sets

$$\Omega^{\mathcal{A}^{(N,S)}}_{L,T} := \{(\vec{l}, \vec{j}) : \vec{l} \in \mathbb{N}_0^{d\cdot N},\ \vec{j} \in \mathbb{Z}^{d\cdot N},\ \lambda_{\text{mix}}(\vec{l}) \cdot \lambda_{\text{iso}}(\vec{l})^{-T} \le (2^L)^{1-T},\ (\boldsymbol{l}_1, \boldsymbol{j}_1) < \cdots < (\boldsymbol{l}_S, \boldsymbol{j}_S) \text{ and } (\boldsymbol{l}_{S+1}, \boldsymbol{j}_{S+1}) < \cdots < (\boldsymbol{l}_N, \boldsymbol{j}_N)\}.$$

Obviously, the inclusions $V^{\mathcal{A}^{(N,S)}}_{K,T_1} \subset V^{\mathcal{A}^{(N,S)}}_{K,T_2}$ for $T_1 \le T_2$ hold. Note that for the associated error the same type of estimate as in Lemma 1 holds. The number of $\vec{l}$-subbands however, i.e. the number of subsets of indices from $\Omega^{\mathcal{A}^{(N,S)}}_{L,T}$ with the same $\vec{l}$, is reduced by the factor $S!(N-S)!$.

## 6. Regularity and decay properties of the solution

So far we introduced various semidiscrete sparse grid spaces for particle problems and carried these techniques over to the case of antisymmetric wave functions. Here, the order of the error estimate depended on the degree $s$ of the Sobolev-norm in which we measure the approximation error and the degrees $t$ and $r$ of anisotropic and isotropic smoothness, respectively, which was assumed to hold for the continuous wave function.

We now return to the electronic Schrödinger problem (1) and invoke our general theory for this special case. To this end, let us recall a major result from [67]. There, Yserentant showed with the help of Fourier transforms that an antisymmetric solution of the electronic Schrödinger equation with $d = 3$ possesses $\mathcal{H}^{1,1}_{\text{mix}}$-regularity in the case $S = 0$ or $S = N$ and at least $\mathcal{H}^{1/2,1}_{\text{mix}}$-regularity otherwise. The main argument to derive this fact is a Hardy type inequality, see [67] for details.

Let us first consider the case of a full antisymmetric solution, i.e. the case $S = 0$ or $S = N$, and the resulting approximation rate in more detail. If we measure the approximation error in the $\mathcal{H}^1$-norm, we obtain from Lemma 1 with $s = 1$ and $t = r = 1$ the approximation order $O((2^L)^{-1+T\cdot\frac{N-1}{N-T}})$ for $T \geq 0$ and $O(2^{-L})$ for $T \leq 0$. In particular, for the choice $T = 0$ we have a rate of $O(2^{-L})$. Also note that the constant in the estimate still depends on $N$ and $d$.

In an analog way we can argue for the partial antisymmetric case where we have for an arbitrary chosen $1 \leq S \leq N$ at least $\mathcal{H}^{1/2,1}_{\text{mix}}$-regularity of the associated wave function. If we measure the approximation error in the $\mathcal{H}^1$-norm, we obtain from Lemma 1 with $s = 1$ and $t = 1/2$, $r = 1$ ($\mathcal{H}^{1/2,1}_{\text{mix}}$-regularity) the approximation order $O((2^{L/2})^{-1+T\cdot\frac{N-1}{N-T}})$ for $T \geq 0$ and $O(2^{-L/2})$ for $T \leq 0$. In particular, for the choice $T = 0$ we have a rate of $O(2^{-L/2})$.

Note however that the order constant depends on $N$ and $d$. Moreover, also the $\mathcal{H}^{1,1}_{\text{mix}}$- and $\mathcal{H}^{1/2,1}_{\text{mix}}$-terms may grow exponentially with the number $N$ of electrons. This is a serious problem for any further discretization in $\vec{j}$-space since to compensate for this exponential growth, the parameter $L$ has to be chosen dependent on $N$. Such a behavior could be seen in the case of a finite domain with periodic boundary conditions with Fourier bases from the results of the numerical experiments in [30] and was one reason why problems with higher numbers of electrons could not be treated.

In [69], a rescaling of the mixed Sobolev norm is suggested. To this end, a scaled analog of the $\mathcal{H}^{1,r}_{\text{mix}}$-norm, $r \in \{0, 1\}$, albeit in Fourier space notation (one $\vec{k}$-scale in Fourier space only instead of the $\vec{l}$- and $\vec{j}$-scales in wavelet space) is introduced, compare also [30], via

$$\|\Psi\|_{\mathcal{H}^{1,r}_{\text{mix}}} = \int_{\mathbb{R}^{dN}} \Big(\prod_{p\in I}\Big(1+\Big|\frac{\boldsymbol{k}_p}{R}\Big|^2\Big)\Big)\Big(\sum_{p=1}^{N}\Big|\frac{\boldsymbol{k}_p}{R}\Big|^2\Big)^r |\hat{\Psi}(\vec{\boldsymbol{k}})|^2 d\vec{\boldsymbol{k}} \tag{28}$$

where $I$ denotes the subset of indices of electrons with the same spin, $\hat{\Psi}(\vec{\boldsymbol{k}})$ is the

Fourier transform of $\Psi$ and $\vec{\boldsymbol{k}} \in \mathbb{Z}^{dN}$ are the coordinates in Fourier space with single-particle-components $\boldsymbol{k}_p \in \mathbb{R}^d$. Here the scaling parameter $R$ relates to the intrinsic length scale of the atom or molecule under consideration. It must hold $R \leq C\sqrt{N}\max(N, Z)$ with $Z = \sum_q Z_q$ the totals charge of the nuclei, see also [56], [69]. For an electronically neutral system $Z = N$ and thus $R \leq CN^{3/2}$. Compared to our definitions $\lambda_{\text{mix}}$ and $\lambda_{\text{iso}}$ of (18) we see the following difference: Besides that (28) involves integration instead of summation, (28) deals with the non-octavized case whereas we used the octavized version which involves powers of two. This is one reason why in the product $\prod_{p\in I}(1 + |\boldsymbol{k}_p/R|^2)$ the factor one must be present. Otherwise the case $\boldsymbol{k}_p = \boldsymbol{0}$ is not dealt properly with. But this also opens the possibility to treat the coordinates with values zero differently in the scaling, since the scaling with $1/R$ in the product acts only on the coordinates with non-zero values.

Furthermore, with $I_+$ and $I_-$ the sets of indices $p$ of electrons for which the spin attains the values $-1/2$ and $1/2$, respectively, and a parameter $K$ (non-octavized case), the subdomain

$$H^Y_{R,K} := \left\{(\boldsymbol{k}_1, \ldots \boldsymbol{k}_N) \in (\mathbb{R}^3)^N : \prod_{p\in I_+}\left(1 + \left|\frac{\boldsymbol{k}_p}{R}\right|_2^2\right) + \prod_{p\in I_-}\left(1 + \left|\frac{\boldsymbol{k}_p}{R}\right|_2^2\right) \leq K^2\right\} \tag{29}$$

in Fourier space describes a cartesian product of two scaled hyperbolic crosses. In the extreme cases $S = 0$ or $S = N$ it degenerates to just one hyperbolic cross. Then, with the projection

$$(P_{R,K}\Psi)(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^{3N} \int \hat{\chi}_{R,K}(\vec{\boldsymbol{k}})\hat{\Psi}(\vec{\boldsymbol{k}})e^{i\vec{\boldsymbol{k}}\cdot\vec{\boldsymbol{x}}}d\vec{\boldsymbol{k}},$$

where $\hat{\chi}_{R,K}$ is the characteristic function of the domain $H^Y_{R,K}$, the following error estimate is shown in [69]: For all eigenfunctions with negative eigenvalues and $s = 0, 1$ there holds

$$\|\Psi - P_{R,K}\Psi\|_s \leq \frac{2\sqrt{e}}{K} R^s \|\Psi\|_0. \tag{30}$$

The restriction to eigenfunctions of the Schrödinger–Hamiltonian whose associated eigenvalues are strictly smaller than zero is not a severe issue since such an assumption holds for bounded states, i.e. any system with localized electrons, compare also [25], [38], [59].

This surprising result shows that, with proper scaling in the norms and the associated choice of a scaled hyperbolic cross, it is possible to get rid of the $\|\Psi\|_{\mathcal{H}^{1,r}_{\text{mix}}}$-terms on the right hand side of sparse grid estimates of the type (22). Note that these terms may grow exponentially with $N$ whereas $\|\Psi\|_0 = 1$. To derive semidiscrete approximation spaces which, e.g. after scaling, overcome this problem is an important step towards any efficient discretization for problems with higher numbers of electrons $N$. Note however that e.g. already for the most simple case $S = 0$ or $S = N$ where in (29) only one cross is involved due to $I_- = \{\,\}$ or $I_+ = \{\,\}$, the subdomain $H^Y_{R,K}$ is no longer a conventional hyperbolic cross in Fourier space. Now, depending of

the different dimensions, the "rays" of the cross are chopped off due to the rescaling with $R$. This gets more transparent if we use the relation

$$\prod_{p=1}^{N}(1+|\boldsymbol{k}_p|_2^2) = \sum_{p=0}^{N}\sum_{\substack{a\subset\{1,\ldots,N\}\\|a|=p}}\prod_{j\in\alpha}|\boldsymbol{k}_j|_2^2$$

and rewrite (29) e.g. in the case $S=0$ or $S=N$ as

$$H_{R,K}^{Y} = \big\{\vec{\boldsymbol{k}} \,:\, K^{-2}\big(\textstyle\sum_{p=0}^{N}\sum_{\substack{a\subset\{1,\ldots,N\}\\|a|=p}} R^{-2p}\prod_{j\in\alpha}|\boldsymbol{k}_j|_2^2\big) \le 1\big\}. \tag{31}$$

If we now define for $K_0, K_1, \ldots, K_N \in \mathbb{N}$

$$H_{K_0,K_1,\ldots,K_N} := \big\{\vec{\boldsymbol{k}} \,:\, \big(\textstyle\sum_{p=0}^{N}\sum_{\substack{a\subset\{1,\ldots,N\}\\|a|=p}} K_p^{-2}\prod_{j\in\alpha}|\boldsymbol{k}_j|_2^2\big) \le 1\big\} \tag{32}$$

we have

$$H_{R^0K,R^1K,\ldots,R^NK} = H_{R,K}^{Y}$$

and see more clearly how the scaling with $R$ acts individually on the different dimensional subsets of the Fourier coordinates. In Figure 4 we give in logarithmic and absolute representation the boundaries of the domains $H_{1,K}^{Y}$, $H_{R,K}^{Y}$ and $H_{1,R^NK}^{Y}$ for $R=8$, $K=2^8$. Here we can observe how the scaled variant $H_{R,K}^{Y}$ is just embedded between the two non-scaled domains $H_{1,K}^{Y}$ and $H_{1,R^NK}^{Y}$. While the boundary of $H_{R,K}^{Y}$ matches in "diagonal" direction that of the huge regular sparse grid $H_{1,R^NK}^{Y}$ this is no longer the case for the other directions.

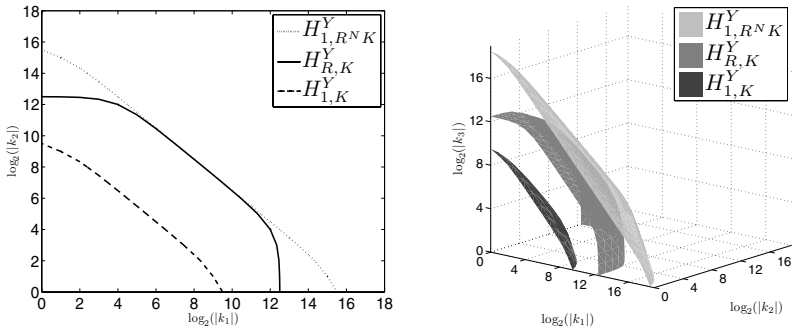



Figure 4. Sets of level indices for $H_{1,K}^{Y}$, $H_{R,K}^{Y}$, $H_{1,R^NK}^{Y}$ in the case $d=1$, $R=8$, $K=2^8$ for $N=2$ and $N=3$.

This dimensional scaling is closely related to well-known decay properties of the solution of Schrödinger's equation which we now recall from the literature. In the seminal work of Agmon [2] the $\mathcal{L}^2$-decay of the eigenfunctions of the electronic

Schrödinger–Hamiltonian of an atom with one nucleus fixed in the origin of the coordinate system is studied in detail and a characterization of the type

$$\int_{\mathbb{R}^{N\cdot d}} |\Psi(\vec{\boldsymbol{x}})|^2 e^{2(1-\varepsilon)\rho(\vec{\boldsymbol{x}})} d\vec{\boldsymbol{x}} \leq c < \infty$$

for any $\varepsilon > 0$ is given for eigenfunctions $\Psi$ with associated eigenvalue $\mu$ below the so-called essential spectrum of $H$. In other words, $\Psi$ decays in the $\mathcal{L}^2$-sense roughly like $e^{-\rho(\vec{\boldsymbol{x}})}$. Here, $\rho(\vec{\boldsymbol{x}})$ is the geodesic distance from $\vec{\boldsymbol{x}}$ to the origin in the Riemannian metric

$$d\vec{\boldsymbol{s}}^2 = (\Lambda_{I(\vec{\boldsymbol{x}})} - \mu) \sum_{i=1}^{N} 2|d\boldsymbol{x}_i|_2^2.$$

To this end, if $I$ denotes any proper subset of $\{1, \ldots N\}$, let $H_I$ denote the restriction of the full Hamiltonian $H$ to the subsystem involving only the electrons associated to $I$ and $\Lambda_I = \inf \sigma(H_I)$, $\Lambda_I = 0$ if $I$ is empty. For any $\vec{\boldsymbol{x}} \in \mathbb{R}^{N\cdot d}/\{0\}$, $I(\vec{\boldsymbol{x}})$ denotes the subset of integers $i \in \{1, \ldots, N\}$ for which $\boldsymbol{x}_i = \boldsymbol{0}$. Note that $\rho$ is *not* isotropic but takes at each point $\vec{\boldsymbol{x}}$ the amount of electrons with position $\boldsymbol{0}$, i.e. the number of electron-nucleus cusps into account.

The result (30) gives some hope that it might indeed be possible to find after additional discretization (in $\vec{\boldsymbol{j}}$-space) an overall discretization which is cost effective and results in an error which does not grow exponentially with the amount of electrons.

The idea is now to decompose the scaled hyperbolic cross $H^Y_{R,K}$ in Fourier space and to approximate the corresponding parts of the associated projection $P_{R,K}\hat{\Psi}(\vec{\boldsymbol{k}})$ properly. To this end, let us assume that we consider a non-periodic, isolated system. It then can be shown that any eigenfunction $\Psi$ with negative eigenvalue below the essential spectrum of the Schrödinger operator $H$ decays exponentially in the $\mathcal{L}^2$-sense with $|\vec{\boldsymbol{x}}| \to \infty$. The same holds for its first derivative [37]. A consequence is that the Fourier transform $\hat{\Psi}$ is infinitely often differentiable as a function in $\vec{\boldsymbol{k}}$. Let us now decompose $H^Y_{R,K}$ into finitely many subdomains $H^Y_{R,K,\vec{\boldsymbol{l}}}$ and let us split $\hat{\Psi}_{R,K}(\vec{\boldsymbol{k}}) := \hat{\chi}_{R,K}(\vec{\boldsymbol{k}})\hat{\Psi}(\vec{\boldsymbol{k}})$ accordingly into $\hat{\Psi}_{R,K,\vec{\boldsymbol{l}}}(\vec{\boldsymbol{k}})$, i.e.

$$\hat{\Psi}_{R,K}(\vec{\boldsymbol{k}}) = \sum_{\vec{\boldsymbol{l}}} \hat{\chi}_{\vec{\boldsymbol{l}}} \hat{\Psi}_{R,K}(\vec{\boldsymbol{k}}) = \sum_{\vec{\boldsymbol{l}}} \hat{\Psi}_{R,K,\vec{\boldsymbol{l}}}(\vec{\boldsymbol{k}})$$

by means of a $C^\infty$-partition of unity $\sum_{\vec{\boldsymbol{l}}} \hat{\chi}_{\vec{\boldsymbol{l}}} = 1$ on $H^Y_{R,K}$, i.e. each $\hat{\chi}_{\vec{\boldsymbol{l}}}(\vec{\boldsymbol{k}}) \in C^\infty$ as a function in $\vec{\boldsymbol{k}}$. Then the functions $\hat{\Psi}_{R,K,\vec{\boldsymbol{l}}}(\vec{\boldsymbol{k}})$ inherit the $C^\infty$-smoothness property and thus can each be well and efficiently approximated by e.g. a properly truncated Fourier series expansion. Note that the detailed choice of partition of unity is not yet specified and there are many possibilities. In the following we will use, e.g. after proper scaling, cf. (10) and (11), the partition

$$\hat{\chi}_{\vec{\boldsymbol{l}}}(\vec{\boldsymbol{k}}) = \prod_{p=1}^{N} \prod_{i=1}^{d} \hat{\chi}_{I_{p,(i)}}(\boldsymbol{k}_{p,(i)})$$

where

$$\hat{\chi}_l(k) := \begin{cases} \hat{\chi}(\frac{k}{c}) & \text{for } l = 0, \\ \hat{\chi}(\frac{k}{c2^l}) - \hat{\chi}(\frac{k}{c2^{l-1}}) & \text{for } l > 0, \end{cases}$$

with

$$\hat{\chi}(k) = \begin{cases} 1 & \text{for } |k| \le \frac{2\pi}{3}, \\ \cos\left( \frac{\pi}{2} \frac{e^{-\frac{4\pi^2}{(3k+2\pi)^2}}}{e^{-\frac{4\pi^2}{(4\pi+3k)^2}} + e^{-\frac{4\pi^2}{(3x+2\pi)^2}}} \right)^2 & \text{for } \frac{2\pi}{3} \le |k| \le \frac{4\pi}{3}, \\ 0 & \text{for } |k| \ge \frac{4\pi}{3}. \end{cases}$$

This choice results in just a representation with respect to the Meyer wavelet series with $\nu^\infty$, i.e. (9) with $\alpha = 2$, compare also (7). The Fourier series expansion of each $\hat{\Psi}_{R,K,\vec{l}}(\vec{k})$ then introduces just the $\vec{j}$-scale, while the $\vec{k}$-scale of the Fourier space relates to the $\vec{l}$-scale of the Meyer wavelets. All we now need is a good decomposition of $H^Y_{R,K}$ into subdomains, a choice of smooth $\hat{\chi}_{\vec{l}}$'s and a proper truncation of the Fourier series expansion of each of the $\hat{\Psi}_{R,K,\vec{l}}$'s. This corresponds to a *truncation* of the Meyer wavelet expansion of $\Psi$ in $\mathbb{R}^{d\cdot N}$ with respect to both the $\vec{l}$- and the $\vec{j}$-scale. Presently, however, it is not completely clear what choice of decomposition and what kind of truncation of the expansion within each subband $\vec{l}$ is most favourable with respect to both the resulting number $M$ of degrees of freedom and the corresponding accuracy of approximation for varying number $N$ of electrons. Anyway, with the choice $K = 2^L$ the set of indices in $\vec{l}, \vec{j}$-wavelet space which is associated to (29) reads

$$\Omega^{\mathcal{A}^{(N,S)}}_{H^Y_{R,2^L}} := \Big\{ (\vec{l}, \vec{j}) \in \Omega^{\mathcal{A}^{(N,S)}} \ : \ \prod_{i=1}^{S} \left(1 + \left|\frac{\tilde{\lambda}(\boldsymbol{l}_i)}{R}\right|_2^2\right) + \prod_{i=S+1}^{N} \left(1 + \left|\frac{\tilde{\lambda}(\boldsymbol{l}_i)}{R}\right|_2^2\right) \le 2^{2L} \Big\},$$

where for $\boldsymbol{l} \in \mathbb{N}_0^d$ we define

$$\tilde{\lambda}(\boldsymbol{l}) := \min_{\boldsymbol{k} \in \mathrm{supp}(\hat{\chi}_{\boldsymbol{l}})} \{|\boldsymbol{k}|_2\}.$$

Note that this involves a kind of octavization due to the size of the support of the $\hat{\chi}_{\boldsymbol{l}}$. For example, we obtain for the Shannon wavelet $\tilde{\lambda}(\boldsymbol{l}) = |(\tilde{\lambda}_{\nu^0}(l_1), \ldots, \tilde{\lambda}_{\nu^0}(l_d))|_2$ with

$$\tilde{\lambda}_{\nu^0}(l) = \begin{cases} 0 & \text{for } l = 0, \\ c\pi 2^{l-1} & \text{otherwise.} \end{cases}$$

## 7. Numerical experiments

We now consider the assembly of the discrete system matrix which is associated to a generalized antisymmetric sparse grid space $V_\Lambda^{\mathcal{A}^{(N,S)}}$ with corresponding finite-dimensional set $\Omega_\Lambda^{\mathcal{A}^{(N,S)}} \subset \Omega^{\mathcal{A}^{(N,S)}}$ and basis functions $\{\Phi_{\vec{l},\vec{j}}^{(N,S)} : (\vec{l},\vec{j}) \in \Omega_\Lambda^{\mathcal{A}^{(N,S)}}\}$ with $\{\Phi_{\vec{l},\vec{j}}^{(N,S)}\}$ from (25) in a Galerkin discretization of (1). To this end, we fix $N > 0$ and $0 \le S \le N$ and omit for reasons of simplicity the indices $S$ and $N$ in the following.

To each pair of indices $(\vec{l},\vec{j})$, $(\vec{l}',\vec{j}')$, each from $\Omega_\Lambda^{\mathcal{A}^{(N,S)}}$, and associated functions $\Phi_{\vec{l},\vec{j}}^{(N,S)}$, $\Phi_{\vec{l}',\vec{j}'}^{(N,S)}$ we obtain one entry in the stiffness matrix, i.e.

$$A_{(\vec{l},\vec{j}),(\vec{l}',\vec{j}')} := \langle \Phi_{\vec{l},\vec{j}}^{(N,S)} | H | \Phi_{\vec{l}',\vec{j}'}^{(N,S)} \rangle = \int \Phi_{\vec{l},\vec{j}}^{(N,S)*}(\vec{x}) H \Phi_{\vec{l}',\vec{j}'}^{(N,S)}(\vec{x})\, \mathrm{d}\vec{x}. \tag{33}$$

Since we use $\mathcal{L}^2$-orthogonal one-dimensional Meyer wavelets as basic building blocks in our construction, also the one-particle basis functions are $\mathcal{L}^2$-orthogonal and we furthermore have $\mathcal{L}^2$-orthogonality of the antisymmetric many-particle basis functions $\Phi_{\vec{l},\vec{j}}^{(N,S)}(\boldsymbol{x})$. We then can take advantage of the well-known Slater–Condon rules [18], [55], [60]. Consequently, quite a few entries of the system matrix are zero and the remaining non-zero entries can be put together from the values of certain $d$- and $2d$-dimensional integrals. These integrals can be written in terms of the Fourier transformation of the Meyer wavelets. In case of the kinetic energy operator we obtain for $\boldsymbol{l}_\alpha, \boldsymbol{l}_\beta \in \mathbb{N}_0^d$ and $\boldsymbol{j}_\alpha, \boldsymbol{j}_\beta \in \mathbb{Z}^d$

$$\begin{aligned}\langle \psi_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha} | -\frac{1}{2}\Delta | \psi_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta} \rangle &= \frac{1}{2}\int_{\mathbb{R}^d} \nabla \psi^*_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha}(\boldsymbol{x}) \cdot \nabla \psi_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\sum_{\mu=1}^{d} \int_{\mathbb{R}} k_\mu^2 \hat{\psi}^*_{\boldsymbol{l}_{\alpha,(\mu)},\boldsymbol{j}_{\alpha,(\mu)}}(k_\mu) \hat{\psi}_{\boldsymbol{l}_{\beta,(\mu)},\boldsymbol{j}_{\beta,(\mu)}}(k_\mu)\, dk_\mu \prod_{\nu\neq\mu}^{d} \delta_{\boldsymbol{l}_{\alpha,(\nu)},\boldsymbol{l}_{\beta,(\nu)}} \delta_{\boldsymbol{j}_{\alpha,(\nu)},\boldsymbol{j}_{\beta,(\nu)}}\end{aligned}$$

and for the integrals related to the $d$-dimensional Coulomb operator $v(\boldsymbol{x}) = 1/|\boldsymbol{x}|_2$ we can write

$$\begin{aligned}\langle \psi_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha} | v | \psi_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta} \rangle &= \int_{\mathbb{R}^d} \psi^*_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha}(\boldsymbol{x}) v(\boldsymbol{x}) \psi_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \hat{v}(\boldsymbol{k}) (\hat{\psi}_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha} * \hat{\psi}_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta})(\boldsymbol{k})\, d\boldsymbol{k}.\end{aligned}$$

For $\boldsymbol{l}_\alpha, \boldsymbol{l}_\beta, \boldsymbol{l}_{\alpha'}, \boldsymbol{l}_{\beta'} \in \mathbb{N}_0^d$ and $\boldsymbol{j}_\alpha, \boldsymbol{j}_\beta, \boldsymbol{j}_{\alpha'}, \boldsymbol{j}_{\beta'} \in \mathbb{Z}^d$ we obtain the integrals related to the electron-electron operator $v(\boldsymbol{x}-\boldsymbol{y}) = 1/|\boldsymbol{x}-\boldsymbol{y}|_2$ in the form

$$\begin{aligned}&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi^*_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha}(\boldsymbol{x}) \psi^*_{\boldsymbol{l}'_\alpha,\boldsymbol{j}_{\alpha'}}(\boldsymbol{y}) v(\boldsymbol{x}-\boldsymbol{y}) \psi_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta}(\boldsymbol{x}) \psi_{\boldsymbol{l}_{\beta'},\boldsymbol{j}_{\beta'}}(\boldsymbol{y}) d\boldsymbol{x}\, d\boldsymbol{y} \\ &\qquad = (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \hat{v}(\boldsymbol{k}) (\hat{\psi}_{\boldsymbol{l}_\alpha,\boldsymbol{j}_\alpha} * \hat{\psi}_{\boldsymbol{l}_\beta,\boldsymbol{j}_\beta})(\boldsymbol{k}) (\hat{\psi}_{\boldsymbol{l}_{\alpha'},\boldsymbol{j}_{\alpha'}} * \hat{\psi}_{\boldsymbol{l}_{\beta'},\boldsymbol{j}_{\beta'}})(\boldsymbol{k})\, d\boldsymbol{k}.\end{aligned}$$

Here, $f * g$ denotes the Fourier convolution, namely $(2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} f(\boldsymbol{x} - \boldsymbol{y}) g(\boldsymbol{y})\, d\boldsymbol{y}$. Note that, in the case of the Meyer wavelet tensor-product basis, the $d$-dimensional Fourier convolution can be written in terms of the one-dimensional Fourier convolution

$$(\hat{\psi}_{\boldsymbol{l}_\alpha, \boldsymbol{j}_\alpha} * \hat{\psi}_{\boldsymbol{l}_\beta, \boldsymbol{j}_\beta})(\boldsymbol{k}) = \prod_{\mu=1}^{d} (\hat{\psi}_{l_{\alpha,(\mu)}, j_{\alpha,(\mu)}} * \hat{\psi}_{l_{\beta,(\mu)}, j_{\beta,(\mu)}})(k_\mu).$$

Thus the $d$-dimensional and $2d$-dimensional integrals in real space which are associated to the Coulomb operator and the electron-electron operator can be written in form of $d$-dimensional integrals of terms involving one-dimensional convolution integrals.

For the solution of the resulting discrete eigenvalue problem we invoke a parallelized conventional Lanczos method taken from the software package SLEPc [35] which is based on the parallel software package PETSc [6]. Note that here also other solution approaches are possible with improved complexities, like multigrid-type methods [13], [15], [44], [47] which however still need to be carried over to the setting of our generalized antisymmetric sparse grids.

Note that an estimate for the accuracy of an eigenfunction relates to an analogous estimate for the eigenvalue by means of the relation $|E - E^{\text{app}}| \leq 4 \cdot \|\Psi - \Psi^{\text{app}}\|^2_{\mathcal{L}^2}$ where $E$ and $\Psi$ denote the exact minimal eigenvalue and associated eigenfunction of $H$, respectively, and $E^{\text{app}}$ and $\Psi^{\text{app}}$ denote finite-dimensional Galerkin approximations in arbitrary subspaces, see also [66].

Then, with Lemma 1, we would obtain for the case $d = 3$ with $s = 0$ and, for example, $r = 1$, $t = 1$ and $S = 0$ the estimate

$$\left|E - E^{\mathcal{A}^{(N,0)}}_{L,T}\right| \leq 4 \cdot \left\|\Psi - \Psi^{\mathcal{A}^{(N,0)}}_{L,T}\right\|^2_{\mathcal{L}^2} \leq O((2^L)^{2\cdot(-2+(T+1)\frac{N-1}{N-T})}) \cdot \left\|\Psi^{\mathcal{A}^{(N,0)}}\right\|^2_{\mathcal{H}^{1,1}_{\text{mix}}}$$

and we see that the eigenvalues are in general much better approximated than the eigenfunctions. For example, for $T = 0$, this would result in a (squared) rate of the order $-4 + 2(N-1)/N$ which is about $-4$ for small numbers of $N$ but gets $-2$ for $N \to \infty$.

Let us now describe our heuristic approach for a finite-dimensional subspace choice in wavelet space which hopefully gives us efficient a-priori patterns $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda}$ and associated subspaces $V^{\mathcal{A}^{(N,S)}}_{\Lambda}$. We use a model function of the Hylleraas-type [16], [41], [57][4]

$$h(\vec{\boldsymbol{x}}) = \prod_{p=1}^{N} \left( e^{-\alpha_p |\boldsymbol{x}_p|_2} \prod_{q>p}^{N} e^{-\beta_{p,q} |\boldsymbol{x}_p - \boldsymbol{x}_q|_2} \right) \tag{34}$$

which reflects the decay properties, the nucleus cusp and the electron-electron cusps of an atom in real space with nucleus fixed in the origin as guidance to a-priori derive a pattern of active wavelet indices in space and scale similar to the simple

[4]Note that we omitted here any prefactors for reasons of simplicity.

one-dimensional example of Figure 2. The localization peak of a Meyer wavelet $\psi_{l,j}$ in real space (e.g. after proper scaling with some $c$ analogously to (10)) is given by

$$\theta(l, j) = \iota(l, j)2^{-l} \quad \text{where } \iota(l, j) = \begin{cases} j & \text{for } l = 0, \\ 1 + 2j & \text{otherwise,} \end{cases}$$

which leads in the multidimensional case to

$$\theta_{\vec{l},\vec{j}} = (\theta(l_1, j_1), \ldots, \theta(l_d, j_d)) \in \mathbb{R}^d$$
$$\theta_{\vec{l},\vec{j}} = (\theta(\boldsymbol{l}_1, \boldsymbol{j}_1), \ldots, \theta(\boldsymbol{l}_N, \boldsymbol{j}_N)) \in (\mathbb{R}^d)^N$$

We now are in the position to describe different discretizations with respect to both the $\vec{\boldsymbol{l}}$-scale and the $\vec{\boldsymbol{j}}$-scale. We focus with respect to the $\vec{\boldsymbol{j}}$-scale on three cases: First, we restrict the whole real space to a finite domain and take the associated wavelets on all incorporated levels into account. Note that in this case the number of wavelets grows from level to level by a factor of 2. Second, we use on each level the same prescribed fixed number of wavelets. And third we let the number of wavelets decay from level to level by a certain factor which results in a multivariate analog to the triangular subspace of Figure 2 (right). With respect to the $\vec{\boldsymbol{l}}$-scale we rely in all cases on the regular sparse grid with $T = 0$. These three different discretization approaches are illustrated in the Figures 5–8 for $d = 1$, $N = 1$ and $d = 1$, $N = 2$, respectively.

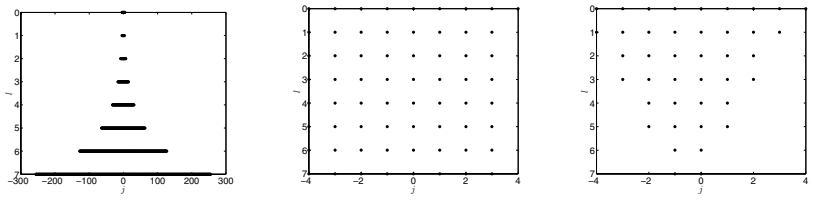


Figure 5. From left to right: Index sets $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ with $d = 1$, $N = 1$, $L = 8$, $J = 4$, $R = 1$ and $\alpha_1 = 1$.

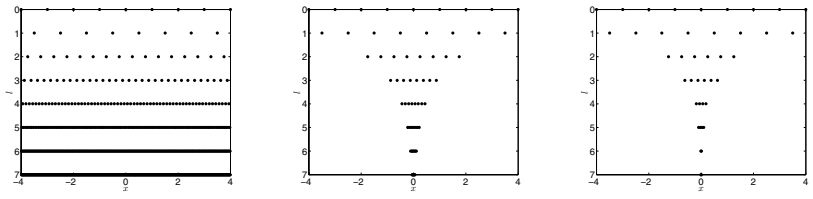


Figure 6. From left to right: Localization peaks of basis functions in real space corresponding to index sets $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ with $d = 1$, $N = 1$, $L = 8$, $J = 4$, $R = 1$ and $\alpha_1 = 1$.

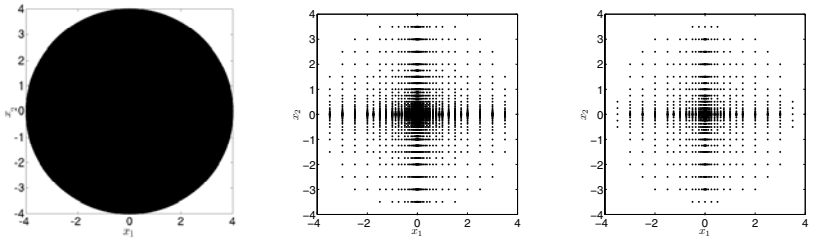


Figure 7. From left to right: Localization peaks of basis functions in real space corresponding to the index sets $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ with $d = 2$, $N = 1$, $L = 8$, $J = 4$, $R = 1$ and $\alpha_1 = 1$.

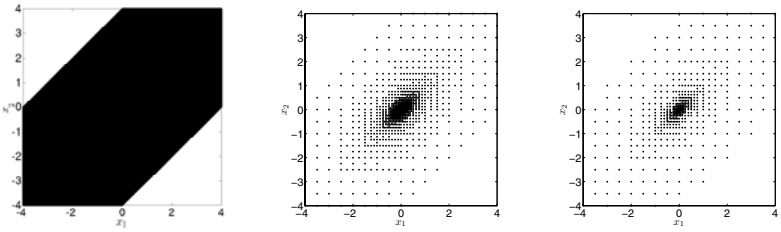


Figure 8. From left to right: Localization peaks of basis functions in real space corresponding to the index sets $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ with $d = 1$, $N = 2$, $L = 8$, $J = 4$, $R = 1$, $S = 1$ and $\alpha_1 = \alpha_2 = \beta_{1,2} = \frac{1}{2}$.

To this end, we define with the parameter $J \in \mathbb{N}_+$ the pattern for the finite domain with full wavelet resolution, i.e. the *full* space (with respect to $\vec{\boldsymbol{j}}$-scale after a finite domain is fixed), as

$$\begin{aligned}\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)} &:= \left\{(\vec{\boldsymbol{l}}, \vec{\boldsymbol{j}}) \in \Omega^{\mathcal{A}^{(N,S)}}_{H^Y_{R,2^L}} : h(\theta(\vec{\boldsymbol{l}}, \vec{\boldsymbol{j}})) > e^{-J}\right\} \\ &= \left\{(\vec{\boldsymbol{l}}, \vec{\boldsymbol{j}}) \in \Omega^{\mathcal{A}^{(N,S)}}_{H^Y_{R,2^L}} : \sum_{p=1}^{N} \left(\alpha_p |\theta(\boldsymbol{l}_p, \boldsymbol{j}_p)|_2 \right.\right. \\ &\qquad \left.\left. + \sum_{q>p}^{N} \beta_{p,q} |\theta(\boldsymbol{l}_p, \boldsymbol{j}_p) - \theta(\boldsymbol{l}_q, \boldsymbol{j}_q)|_2\right) < J\right\}\end{aligned}$$

with prescribed $\alpha_p$, $\beta_{p,q}$. Note here the equivalence of the sum to $\ln(h(\theta(\vec{\boldsymbol{l}}, \vec{\boldsymbol{j}})))$.

To describe the other two cases we set with a general function $\Theta$ which still has to be fixed

$$\begin{aligned}&\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta}(L,J,R)} := \left\{(\vec{\boldsymbol{l}}, \vec{\boldsymbol{j}}) \in \Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)} :\right. \\ &\left.\sum_{p=1}^{N} \left(\alpha_p \left|\Theta\left(\boldsymbol{l}_p, \boldsymbol{j}_p\right)\right|_2 + \sum_{q>p}^{N} \beta_{p,q} \left|\Theta\left(\theta^{-1}\left(\theta\left(\boldsymbol{l}_p, \boldsymbol{j}_p\right) - \theta\left(\boldsymbol{l}_q, \boldsymbol{j}_q\right)\right)\right)\right|_2\right) < J\right\}.\end{aligned}$$

Note that $\theta^{-1}$ denotes the inverse mapping to $\theta$. It holds

$$\theta^{-1}\left(\theta\,(l,\,j)-\theta\left(l',\,j'\right)\right)=\begin{cases}\tilde{\iota}^{-1}(l,\,\iota(l,\,j)-\iota(l',\,j')2^{l-l'}) & \text{for } l\geq l',\\ \tilde{\iota}^{-1}(l',\,\iota(l,\,j)2^{l'-l}-\iota(l',\,j')) & \text{for } l'\geq l\end{cases}$$

where $\tilde{\iota}(l,\,j)=(l,\,\iota(l,\,j))$. We now define the *rectangular* index set $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\mathrm{rec}}(L,J,R)}$ via the following choice of $\Theta$: For $\boldsymbol{l}\in\mathbb{N}_0^d$ and $\boldsymbol{j}\in\mathbb{Z}^d$ we set

$$\Theta_{\mathrm{rec}}(\boldsymbol{l},\,\boldsymbol{j}):=(\Theta_{\mathrm{rec}}(l_1,\,j_1),\ldots,\Theta_{\mathrm{rec}}(l_d,\,j_d))$$

and for $l\in\mathbb{N}_0$ and $j\in\mathbb{Z}$ we set

$$\Theta_{\mathrm{rec}}(l,\,j):=\begin{cases}|j|, & \text{for } l=0,\\ |\frac{1}{2}+j| & \text{otherwise.}\end{cases}$$

Finally we define the *triangle* space $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\mathrm{tri}}(L,J,R)}$ with help of

$$\Theta_{\mathrm{tri}}(\boldsymbol{l},\,\boldsymbol{j}):=(\Theta_{\mathrm{tri}}(l_1,\,j_1),\ldots,\Theta_{\mathrm{tri}}(l_d,\,j_d))$$

where for $l\in\mathbb{N}_0$ and $j\in\mathbb{Z}$ we set

$$\Theta_{\mathrm{tri}}(l,\,j):=\begin{cases}\frac{|j|}{1-\frac{l}{L_{\max}+1}} & \text{for } l=0,\\ \frac{|\frac{1}{2}+j|}{1-\frac{l}{L_{\max}+1}} & \text{for } 0<l\leq L_{\max},\\ \infty & \text{otherwise}\end{cases}$$

with $L_{\max}$ as the maximum level for the respective triangle.

Let us now discuss the results of our first, very preliminary numerical experiments with these new sparse grid methods for Schrödinger's equation. To this end, we restrict ourselves for complexity reasons to the case of one-dimensional particles only. The general three-dimensional case will be the subject of a forthcoming paper. We use in the following in (1) the potential

$$V=-\sum_{p=1}^{N}\sum_{q=1}^{N_{\mathrm{nuc}}}Z_q\,v(\boldsymbol{x}_p-\boldsymbol{R}_q)+\sum_{p=1}^{N}\sum_{q>p}^{N}v(\boldsymbol{x}_p-\boldsymbol{x}_q)\tag{35}$$

with

$$v(\boldsymbol{r})=\begin{cases}D-|\boldsymbol{r}|_2 & \text{for } |\boldsymbol{r}|_2\leq D,\\ 0 & \text{otherwise}\end{cases}$$

which is truncated at radius $D$ and shifted by $D$. Note that $\lim_{|\boldsymbol{r}|_2\to\infty}v(\boldsymbol{r})=0$. Up to truncation and the shift with $D$, $|\boldsymbol{r}|_2$ is just the one-dimensional analogue to the

Coulomb potential. The Fourier transform reads

$$\hat{v}(\boldsymbol{k}) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}} \frac{1}{|\boldsymbol{k}|_2^2}(1 - \cos(D|\boldsymbol{k}|_2)) & \text{for } |\boldsymbol{k}|_2 \neq 0, \\ \frac{D^2}{\sqrt{2\pi}} & \text{otherwise.} \end{cases}$$

Note that $\hat{v}$ is continuous.

We study for varying numbers $N$ of particles the behavior of the discrete energy $E$, i.e. the smallest eigenvalue of the associated system matrix $A$, as $L$ and $J$ increase. Here, we use the generalized antisymmetric sparse grids $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ and focus on the two cases $S = 0$ or $S = \lfloor N/2 \rfloor$. We employ the Meyer wavelets with (9) where $\nu^\infty$, $\alpha = 2$, and the Shannon wavelet with $\nu^0$ from (8). Tables 1 and 2 give the obtained results. Here, $M$ denotes the number of degrees of freedom and $\#A$ denotes the number of the non-zero matrix entries. Furthermore, $\Delta E$ denotes the difference of the obtained values of $E$ and $\varepsilon$ denotes the quotient of the values of $\Delta E$ for two successive rows in the table. Thus, $\varepsilon$ indicates the convergence rate of the discretization error.

Table 1. $d = 1$, $N = 1$, $c = 1$, $R = 1$, $\alpha_1 = 1$, $L_{\max} = L$, $D = 8$.

| $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$ | | | | $\nu^\infty$ | | | $\nu^0$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 2 | 1 | 5 | 25 | −7.187310 | | | −7.186261 | | |
| 4 | 1 | 9 | 81 | −7.189322 | 2.01e−03 | | −7.188615 | 2.35e−03 | |
| 8 | 1 | 17 | 289 | −7.189334 | 1.14e−05 | 175.1 | −7.188674 | 5.92e−05 | 39.7 |
| 16 | 1 | 33 | 1089 | −7.189335 | 1.08e−06 | 10.6 | −7.188683 | 9.25e−06 | 6.4 |
| 32 | 1 | 65 | 4225 | −7.189335 | 4.60e−07 | 2.3 | −7.188684 | 1.29e−06 | 7.1 |
| 64 | 1 | 129 | 16641 | −7.189335 | 1.36e−09 | 336.4 | −7.188685 | 1.73e−07 | 7.4 |
| 16 | 1 | 33 | 1089 | −7.189335 | | | −7.188683 | | |
| 16 | 2 | 65 | 4225 | −7.191345 | 2.00e−03 | | −7.190920 | 2.23e−03 | |
| 16 | 3 | 129 | 16641 | −7.191376 | 3.19e−05 | 62.9 | −7.190958 | 3.80e−05 | 58.7 |
| 16 | 4 | 257 | 66049 | | | | −7.190959 | 1.00e−06 | 37.6 |
| 16 | 5 | 513 | 263169 | | | | −7.190959 | 3.04e−08 | 33.1 |
| $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ | | | | $\nu^\infty$ | | | $\nu^0$ | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 16 | 1 | 33 | 1089 | −7.189335 | | | −7.188683 | | |
| 16 | 2 | 65 | 4225 | −7.191345 | 2.00e−03 | | −7.190920 | 2.23e−03 | |
| 16 | 3 | 97 | 9409 | −7.191376 | 3.19e−05 | 62.9 | −7.190956 | 3.61e−05 | 61.8 |
| 16 | 4 | 129 | 16641 | −7.191377 | 8.37e−07 | 38.0 | −7.190957 | 9.52e−07 | 37.9 |
| 16 | 5 | 161 | 25921 | −7.191377 | 2.51e−08 | 33.3 | −7.190957 | 2.85e−08 | 33.3 |
| 16 | 6 | 193 | 37249 | −7.191377 | 7.53e−10 | 33.4 | −7.190957 | 8.84e−10 | 32.2 |
| $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ | | | | $\nu^\infty$ | | | $\nu^0$ | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 16 | 1 | 33 | 1089 | −7.189335 | | | −7.189335 | | |
| 16 | 2 | 55 | 3025 | −7.191314 | 1.97e−03 | | −7.190747 | 1.41e−03 | |
| 16 | 3 | 73 | 5329 | −7.191357 | 4.25e−05 | 46.5 | −7.190825 | 7.80e−05 | 18.0 |
| 16 | 4 | 91 | 8281 | −7.191366 | 9.28e−06 | 4.5 | −7.190865 | 4.04e−05 | 1.9 |
| 16 | 5 | 107 | 11449 | −7.191366 | 2.13e−07 | 43.4 | −7.190866 | 5.42e−07 | 74.5 |
| 16 | 6 | 125 | 15625 | −7.191371 | 5.07e−06 | 0.042 | −7.190900 | 3.39e−05 | 0.016 |

In Table 1, with just one particle, i.e. $N = 1$, we see that the minimal eigenvalues for the Shannon wavelet are slightly, i.e. by $10^{-3} - 10^{-2}$, worse than the minimal eigenvalues for the Meyer wavelet with $\nu^\infty$. Furthermore, from the first part of the table where we fix $L = 1$ and vary $J$ and alternatively fix $J = 16$ and vary $L$ it gets clear that it necessary to increase both $J$ *and* $L$ to obtain convergence. While just an increase of $J$ with fixed $L = 1$ does not improve the result at all (with $D$ fixed), the increase of $L$ for a fixed $J$ at least gives a convergence to the solution on a bounded domain whose size is associated to the respective value of $D$ and $J$. In the second part of the table we compare the behavior for $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$ and $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ for the wavelets with $\nu^\infty$ and $\nu^0$. While we see relatively stable monotone rates of around 33 and better in case of $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$, the convergence behavior for $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ is more erratic. Nevertheless, when we compare the achieved results for the same amount of matrix entries $\#A$ we see not much difference. For example, with $\nu^\infty$, we get for $J = 16, L = 6$ with 125 degrees of freedom and 15625 matrix entries a value of $-7.191371$ for $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{rec}}}(L,J,R)}$ whereas we get for $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\text{full}}(L,J,R)}$ with $J = 16, L = 4$ with about the same degrees of freedom and matrix entries nearly the same value $-7.191377$.

Let us now consider the results for $N > 1$ given in Table 2. Here we restricted ourself to the sparse grid $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$ due to complexity reasons. We see that the computed minimal eigenvalues in the case $S = \frac{N}{2}$ are higher than that in the case $S = 0$, as to be expected. Furthermore, our results suggest convergence for rising $L$. If we compare the cases $R = 1$ and $R = 2^{\frac{3}{2}}$ for $N = 2$, we see that both the number of degrees of freedom and the minimal eigenvalues are for $R = 1$ approximately the same as for $R = 2^{\frac{3}{2}}$ on the next coarser level. An analogous observation holds in the case $N = 4$.

Note furthermore that the sparse grid effect acts only on the fully antisymmetric subspaces of the total space. This is the reason for the quite large number of degrees of freedom for the case $N = 4$, $S = 2$.

Note finally that our present simple numerical quadrature procedure is relatively expensive. To achieve results for higher numbers of particles with sufficiently large $L$ and $J$, the numerical integration scheme has to be improved. Moreover, to deal in the future with the case of three-dimensional particles using the classical potential (2) and the Meyer wavelets with $\nu^\infty$, an efficient and accurate numerical quadrature still has to be derived.[5]

[5]Such a numerical quadrature scheme must be able to cope with oscillatory functions and also must resolve the singularity in the Coulomb operator.

Table 2. $d = 1$, $c = 1$, $D = 8$, $\Omega^{\mathcal{A}^{(N,S)}}_{\Lambda^{\Theta_{\text{tri}}}(L,J,R)}$, $\nu^0$.

| $Z = 2$, $N = 2$, $S = 1$, $R = 1$, $\alpha_1 = \alpha_2 = \beta_{1,2} = \frac{1}{2}$ | | | | | | |
|---|---|---|---|---|---|---|
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 4 | 1037 | 1029529 | −28.818529 | | |
| 8 | 5 | 1401 | 1788305 | −28.819933 | 1.40e−03 | |
| 8 | 6 | 1623 | 2324081 | −28.819954 | 2.07e−05 | 67.55 |
| 8 | 7 | 1943 | 3240369 | −28.819963 | 8.81e−06 | 2.35 |
| $Z = 2$, $N = 2$, $S = 1$, $R = 2^{\frac{3}{2}}$, $\alpha_1 = \alpha_2 = \beta_{1,2} = \frac{1}{2}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 3 | 1067 | 1092649 | −28.818529 | | |
| 8 | 4 | 1425 | 1856129 | −28.819933 | 1.40e−03 | |
| 8 | 5 | 1637 | 2369721 | −28.819954 | 2.07e−05 | 67.55 |
| 8 | 6 | 1957 | 3829849 | −28.819963 | 8.81e−06 | 2.35 |
| $Z = 2$, $N = 2$, $S = 0$, $R = 1$, $\alpha_1 = \alpha_2 = \beta_{1,2} = \frac{1}{2}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 4 | 383 | 138589 | −27.134075 | | |
| 8 | 5 | 501 | 234965 | −27.134725 | 6.49e−04 | |
| 8 | 6 | 614 | 341696 | −27.134725 | 3.98e−07 | 1630.99 |
| 8 | 7 | 731 | 470569 | −27.134725 | 3.77e−07 | 1.05 |
| $Z = 2$, $N = 2$, $S = 0$, $R = 2^{\frac{3}{2}}$, $\alpha_1 = \alpha_2 = \beta_{1,2} = \frac{1}{2}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 3 | 475 | 215125 | −27.134075 | | |
| 8 | 4 | 622 | 353684 | −27.134725 | 6.49e−04 | |
| 8 | 5 | 714 | 455420 | −27.134725 | 3.98e−07 | 1631.10 |
| 8 | 6 | 852 | 624244 | −27.134725 | 3.77e−07 | 1.05 |
| $Z = 4$, $N = 4$, $S = 2$, $R = 1$, $\alpha_p = \beta_{p,q} = \frac{1}{4}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 4 | 24514 | 17003256 | −106.755154 | | |
| 8 | 5 | 39104 | 32716440 | −106.756364 | 1.20e−03 | |
| $Z = 4$, $N = 4$, $S = 2$, $R = 8$, $\alpha_p = \beta_{p,q} = \frac{1}{4}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 1 | 31592 | 22864800 | −106.755154 | | |
| $Z = 4$, $N = 4$, $S = 0$, $R = 1$, $\alpha_p = \beta_{p,q} = \frac{1}{4}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 4 | 1903 | 313963 | −102.659381 | | |
| 8 | 5 | 2842 | 647688 | −102.659503 | 1.22e−04 | |
| 8 | 6 | 4039 | 1063101 | −102.660489 | 9.86e−04 | 0.12 |
| $Z = 4$, $N = 4$, $S = 0$, $R = 8$, $\alpha_p = \beta_{p,q} = \frac{1}{4}$ | | | | | | |
| $J$ | $L$ | $M$ | $\#A$ | $E$ | $\Delta E$ | $\varepsilon$ |
| 8 | 1 | 3527 | 761851 | −102.659381 | | |
| 8 | 2 | 6029 | 1558219 | −102.659503 | 1.22e−04 | |
| 8 | 3 | 8098 | 2343162 | −102.660489 | 9.85e−04 | 0.12 |

## 8. Concluding remarks

In this article we proposed to use Meyer's wavelets in a sparse grid approach for a direct discretization of the electronic Schrödinger equation. The sparse grid constructions promises to break the curse of dimensionality to some extent and may allow a numerical treatment of the Schrödinger equation without resorting to any model approximation. We discussed the Meyer wavelet family and their properties and built on them an anisotropic multiresolution analysis for general particle spaces. Furthermore

we studied a semidiscretization with respect to the level and introduced generalized semidiscrete sparse grid spaces. We then restricted these spaces to the case of anti-symmetric functions with additional spin. Using regularity and decay properties of the eigenfunctions of the Schrödinger operator we discussed rescaled semidiscrete sparse grid spaces due to Yserentant. They allow to get rid of the terms that involve the $\mathcal{H}^{1,1}_{\text{mix}}$- and $\mathcal{H}^{1/2,1}_{\text{mix}}$-norm of the eigenfunction which may grow exponentially with the number of electrons present in the system. Thus a direct estimation of the approximation error can be achieved that only involves the $\mathcal{L}^2$-norm of the eigenfunction. We also showed that a Fourier series approximation of a splitting of the eigenfunctions living on a scaled hyperbolic cross in Fourier space essentially just results in Meyer wavelets. Therefore, we directly tried to discretize Schrödinger's equation in properly chosen wavelet subspaces.

We only presented preliminary numerical results with one-dimensional particles and a shifted and truncated potential. For the Meyer wavelets with $\nu^\infty$ and for the classical, not truncated Coulomb potential, substantially improved quadrature routines have to be developed in the future to achieved reasonable run times for the set up of the stiffness matrix. Furthermore, the interplay and the optimal choice of the coarsest scale, i.e. of $c$, the scaling parameter $R$, the domain truncation parameter $J$, the scale truncation parameter $L$ and the parameters $L_{\max}$, $\alpha_p$, $\beta_{p,q}$ is not clear at all and needs further investigation. Finally more experiments are necessary with other types of sparse grid subspaces beyond the ones derived from the Hylleraas-type function (34) to complete our search for an accurate and cost effective approximation scheme for higher numbers $N$ of electrons. Probably not the best strategy for subspace selection was yet used and substantially improved schemes can be found in the future. This may be done along the lines of best $M$-term approximation which, from a theoretical point of view, would however involve a new, not yet existing Besov regularity theory for high-dimensional spaces in an anisotropic setting. Or, from a practical point of view, this would involve new adaptive sparse grid schemes using tensor product Meyer wavelets which need proper error estimators and refinement strategies for both the boundary truncation error and, balanced with it, the scale truncation error.

The sparse grid approach is based on a tensor product construction which allows to treat the nucleus–electron cusps properly which are aligned to the particle-coordinate axes of the system but which does not fit to the "diagonal" directions of the electron–electron cusps. Here, proper a-priori refinement or general adaptivity must be used which however involves for $d = 3$ at least the quite costly resolution of three-dimensional manifolds in six-dimensional space which limits the approach. To this end, new features have to brought into the approximation like for example wavelets which allow additionally for multivariate rotations in the spirit of curvelets [14]. Also an approach in the spirit of wave-ray multigrid methods [9] may be envisioned. Alternatively an embedding in still higher-dimensional formulations which allows to express the electron-electron pairs as new coordinate directions might be explored. This, however, is future work.

## References


[1] Adams, R., *Sobolev spaces*. Academic Press, New York 1975.

[2] Agmon, S., *Lectures on exponential decay of solutions of second-order elliptic equations: Bounds on eigenfunctions of N-body Schrödinger operators*. Math. Notes 29, Princeton University Press, Princeton 1982.

[3] Atkins, P., and Friedman, R., *Molecular quantum mechanics*. Oxford University Press, Oxford 1997.

[4] Auscher, P., Weiss, G., and Wickerhauser, G., Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets. In *Wavelets: A tutorial in theory and applications* (ed. by C. K. Chui), Academic Press, New York 1992, 237–256.

[5] Babenko, K., Approximation by trigonometric polynomials in a certain class of periodic functions of several variables. *Dokl. Akad. Nauk SSSR* **132** (1960), 672–675; English transl. *Soviet Math. Dokl.* **1** (1960), 672–675.

[6] Balay, S., Buschelman, K., Eijkhout, V., Gropp, W., Kaushik, D., Knepley, M., McInnes, L., Smith, and Zhang, H., PETSc users manual. Tech. Report ANL-95/11 - Revision 2.1.5, Argonne National Laboratory, 2004.

[7] Balian, R., Un principe d'incertitude fort en théorie du signal on mécanique quantique. *C. R. Acad. Sci. Paris Sér. II* **292** (1981), 1357–1361.

[8] Bellmann, R., *Adaptive control processes: A guided tour*. Princeton University Press, Princeton 1961.

[9] Brandt, A., and Livshits, I., Wave-ray multigrid method for standing wave equations. Electron. *Trans. Numer. Anal.* **6** (1997), 162–181.

[10] Le Bris, C., Computational chemistry from the perspective of numerical analysis. *Acta Numer.* **14** (2005), 363–444.

[11] Bungartz, H., and Griebel, M., A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. *J. Complexity* **15** (1999), 167–199.

[12] Bungartz, H., and Griebel, M., Sparse grids. *Acta Numer.* **13** (2004), 147–269.

[13] Cai, Z., Mandel, J., and McCormick, S., Multigrid methods for nearly singular linear equations and eigenvalue problems. *SIAM J. Numer. Anal.* **34** (1997), 178–200.

[14] Candés, E., and Donoho, D., Ridgelets: a key to higher-dimensional intermittency? *Phil. Trans. Roy. Soc. London Ser. A* **357** (1999), 2495–2509.

[15] Chan, T., and Sharapov, I., Subspace correction multi-level methods for elliptic eigenvalue problems. *Numer. Linear Algebra Appl.* **9** (1) (2002), 1–20.

[16] Chandrasekhar, S., and Herzberg, G., Energies of the ground states of He, $Li^+$, and $O^{6+}$. *Phys. Rev.* **98** (4) (1955), 1050–1054.

[17] Coifman, R., and Meyer, Y., Remarques sur l'analyse de Fourier à fenêtre. *C. R. Acad. Sci. Paris Sér. I Math.* **312** (1991), 259–261.

[18] Condon, E., The theory of complex spectra. *Phys. Rev.* **36** (7) (1930), 1121–1133.

[19] Daubechies, I., *Ten lectures on wavelets*. CBMS-NSF Regional Conf. Series in Appl. Math. 61, SIAM, Philadelphia 1992.

[20] Daubechies, I., Jaffard, S., and Journé, J., A simple Wilson orthonormal basis with exponential decay. *SIAM J. Math. Anal.* **24** (1990), 520–527.

[21] DeVore, R., Konyagin, S., and Temlyakov, V., Hyperbolic wavelet approximation. *Constr. Approx.* **14** (1998), 1–26.

[22] Feynman, R., There's plenty of room at the bottom: An invitation to enter a new world of physics. *Engineering and Science* **XXIII** (Feb. issue) (1960), http://www.zyvex.com/nanotech/feynman.html.

[23] Fliegl, H., Klopper, W., and Hättig, C., Coupled-cluster theory with simplified linear-r12 corrections: The CCSD(R12) model. *J. Chem. Phys.* **122** (8) (2005), 084107.

[24] Frank, K., Heinrich, S., and Pereverzev, S., Information complexity of multivariate Fredholm equations in Sobolev classes. *J. Complexity* **12** (1996), 17–34.

[25] Froese, R., and Herbst, I., Exponential bounds and absence of positive eigenvalues for $N$-body Schrödinger-operators. *Comm. Math. Phys.* **87** (3) (1982), 429–447.

[26] Garcke, J., and Griebel, M., On the computation of the eigenproblems of hydrogen and helium in strong magnetic and electric fields with the sparse grid combination technique. *J. Comput. Phys.* **165** (2) (2000), 694–716.

[27] Gerstner, T., and Griebel, M., Numerical integration using sparse grids. *Numer. Algorithms* **18** (1998), 209–232.

[28] Gerstner, T., and Griebel, M., Dimension–adaptive tensor–product quadrature. *Computing* **71** (1) (2003), 65–87.

[29] Griebel, M., Sparse grids and related approximation schemes for higher dimensional problems. In *Proceedings of the Conference on Foundations of Computational Mathematics* (FoCM05), Santander, Spain, 2005.

[30] Griebel, M., and Hamaekers, J., Sparse grids for the Schrödinger equation. *Math. Model. Numer. Anal.*, submitted.

[31] Griebel, M., and Knapek, S., Optimized tensor-product approximation spaces. *Constr. Approx.* **16** (4) (2000), 525–540.

[32] Griebel, M., Oswald, P., and Schiekofer, T., Sparse grids for boundary integral equations. *Numer. Math.* **83** (2) (1999), 279–312.

[33] Hackbusch, W., The efficient computation of certain determinants arising in the treatment of Schrödinger's equation. *Computing* **67** (2000), 35–56.

[34] Hernández, E., and Weiss, G., *A first course on wavelets*. Stud. Adv. Math., CRC Press, Boca Raton 1996.

[35] Hernandez, V., Roman, J., and Vidal, V., SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Software* **31** (3) (2005), 351–362.

[36] Hochmuth, R., Wavelet bases in numerical analysis and restricted nonlinear approximation. Habilitationsschrift, Freie Universität Berlin, 1999.

[37] Hoffmann-Ostenhof, M., Hoffmann-Ostenhof, T., and Sørensen, T., Electron wavefunction and densities for atoms. *Ann. Henri Poincaré* **2** (2001), 77–100.

[38] Hunziker, W., and Sigal, I., The quantum $N$-body problem. *J. Math. Phys.* **41** (2000), 3448–3510.

[39] Jaffard, S., Meyer, Y., and Ryan, R., *Wavelets: Tools for science and technology*. SIAM, Philadelphia, PA, 2001.

[40] Kaiblinger, N., and Madych, W., Orthonormal sampling functions. *Appl. Comput. Harmon. Anal.*, to appear.

[41] Klopper, W., r12-dependent wavefunctions. In *The Encyclopedia of Computational Chemistry* (ed. by P. von Ragué Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, and P. R. Schreiner), John Wiley and Sons, Chichester 1998, 2351–2375.

[42] Knapek, S., Approximation und Kompression mit Tensorprodukt-Multiskalenräumen. Dissertation, Universität Bonn, 2000.

[43] Knapek, S., Hyperbolic cross approximation of integral operators with smooth kernel. Tech. Report 665, SFB 256, Universität Bonn, 2000.

[44] Knyazev, A., and Neymeyr, K., Efficient solution of symmetric eigenvalue problem using multigrid preconditioners in the locally optimal block conjugate gradient method. *Electron. Trans. Numer. Anal.* **15** (2003), 38–55.

[45] Lemarié, P., and Meyer, Y., Ondelettes et bases hilbertiennes. *Rev. Mat. Iberoamericana* **2** (1–2) (1986), 1–18.

[46] Levine, I., *Quantum chemistry*. 5th ed., Prentice-Hall, 2000.

[47] Livne, O., and Brandt, A., $O(N \log N)$ multilevel calculation of $N$ eigenfunctions. In *Multiscale Computational Methods in Chemistry and Physics* (ed. by A. Brandt, J. Bernholc, and K. Binder), NATO Science Series III: Computer and Systems Sciences 177, IOS Press, 2001, 112–136.

[48] Low, F., Complete sets of wave packets. In *A Passion for Physics—Essays in Honor of Geoffrey Chew*, World Scientific, Singapore 1985, 17–22.

[49] Malvar, H., Lapped transform for efficient transform/subband coding. *IEEE Trans. Acoust. Speech Signal Process.* **38** (1990), 969–978.

[50] Mazziotti, D., Variational two-electron reduced density matrix theory for many-electron atoms and molecules: Implementation of the spin- and symmetry-adapted T-2 condition through first-order semidefinite programming. *Phys. Rev. A* **72** (3) (2005), 032510.

[51] Messiah, A., *Quantum mechanics*. Vol. 1, 2, North-Holland, Amsterdam 1961/62.

[52] Meyer, Y., *Wavelets and operators*. Cambridge Stud. Adv. Math. 37, Cambridge University Press, Cambridge 1992.

[53] Meyer, Y., *Wavelets, vibrations and scalings*. CRM Monograph Ser.9, Amer. Math. Soc., Providence, RI, 1998.

[54] Pan, G., *Wavelets in electromagnetics and device modeling*. Wiley–IEEE Press, 2003.

[55] Parr, R., and Yang, W., *Density functional theory of atoms and molecules*. Oxford University Press, New York 1989.

[56] Persson, A., Bounds for the discrete part of the spectrum of a semi-bounded Schrödinger operator. *Math. Scand.* **8**, (1960), 143–153.

[57] Rodriguez, K., and Gasaneo, G., Accurate Hylleraas-like functions for the He atom with correct cusp conditions. *J. Phys. B: At. Mol. Opt. Phys.* **38** (2005), L259–L267.

[58] Schmeisser, H., and Triebel, H., *Fourier analysis and functions spaces*. John Wiley, Chichester 1987.

[59] Simon, B., Schrödinger operators in the twentieth century. *J. Math. Phys.* **41** (2000), 3523–3555.

[60] Slater, J., The theory of complex spectra. *Phys. Rev.* **34** (10) (1929), 1293–1322.

[61] Smolyak, S., Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR* **148** (1963), 1042–1045; English. transl. *Soviet Math. Dokl.* **4** (1963), 240–243.

[62] Walnut, D., *An introduction to wavelet analysis*. Appl. Numer. Harmon. Anal., Birkhäuser, Boston 2002.

[63] Walter, G., and Zhang, J., Orthonormal wavelets with simple closed-form expressions. *IEEE Trans. Signal Process.* **46** (8) (1998), 2248–2251.

[64] Wilson, K., Renormalization group and critical phenomena. II. Phase-space cell analysis of critical behavior. *Phys. Rev. B* **4** (1971), 3184–3205.

[65] Yamada, M., and Ohkitani, K., An identification of energy cascade in turbulence by orthonormal wavelet analysis. *Prog. Theor. Phys.* **836** (4) (1991), 799–815.

[66] Yserentant, H., On the electronic Schrödinger equation. Report 191, SFB 382, Universität Tübingen, 2003.

[67] Yserentant, H., On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. *Numer. Math.* **98** (2004), 731–759.

[68] Yserentant, H., Sparse grid spaces for the numerical solution of the electronic Schrödinger equation. *Numer. Math.* **101** (2005), 381–389.

[69] Yserentant, H., The hyperbolic cross space approximation of electronic wavefunctions. *Numer. Math.* submitted.

[70] Yserentant, H., The hyperbolic cross space approximation of electronic wavefunctions. Talk at IHP-EU Network Workshop/Winter School Breaking Complexity: Nonlinear/Adaptive Approximation in High Dimensions, Physikzentrum Bad Honnef, Germany, 15th December 2005.

Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany
E-mail: griebel@ins.uni-bonn.de

Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany
E-mail: hamaeker@ins.uni-bonn.de

# Mathematical and numerical analysis for molecular simulation: accomplishments and challenges

Claude Le Bris

**Abstract.** Molecular simulation is explored from the mathematical viewpoint. The field comprises computational chemistry and molecular dynamics. A variety of mathematical and numerical questions raised is reviewed. Placing the models and the techniques employed for simulation on a firm mathematical ground is a difficult task, which has begun decades ago. The time is right for assessing the field, and the issues and challenges ahead.

**Mathematics Subject Classification (2000).** 35Bxx, 35Jxx, 35Pxx, 49Kxx, 81Q05, 81Q10, 82Bxx.

**Keywords.** Computational chemistry, molecular simulation, molecular dynamics, Schrödinger equations, Hartree–Fock theory, Thomas–Fermi theory, Density Functional Theory, nonlinear eigenvalue problems, spectral theory, spectral methods, elliptic partial differential equations, optimization, reduced basis methods, sparse grids, Hamiltonian dynamics, symplectic methods, geometric integration, stochastic differential equations, Markov chains, Monte-Carlo techniques.

## 1. Introduction

*Molecular simulation* is an increasingly important field of scientific computing. It comprises *computational chemistry*, focused on the calculations of electronic structures and the related properties, and *molecular dynamics*, devoted to the simulation of molecular evolutions, evaluations of ensemble averages and thermodynamic quantities. Examples of reference treatises are [68], [70] and [1], [39], respectively. We also refer to [14, Chapter 1] for a mathematically-oriented introductory text.

**1.1. Ubiquity of molecular simulation.** The field has several intimate connections with many other fields. Indeed, *molecular simulation* is above all important because many macroscopic properties of matter originate from phenomena at the microscopic scale. Instances are: electrical conductivities, colors, chemical reactivities, mechanical behaviour, aging. Accurate calculations on representative microscopic systems allow for the evaluation of such properties. Additionally, even the macroscopic phenomena that proceed from bulk effects, and which thus necessitate the consideration of large size microscopic systems, may now be studied by advanced techniques in molecular simulation. Recent record calculations simulate the dynamics of billions

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of atoms over a microsecond. Molecular biology, chemistry and physics are thus inseparable today from molecular simulation. An easy observation sheds some light on this. Roughly one publication out of ten in chemistry journals presents some numerical simulations performed on theoretical models. This is an impressive ratio for a field so much experimentally oriented.

Computations are first seen as complements to experiments. For instance, all the information about the electronic properties is contained in the wave function; the latter cannot be measured but it can be computed. Computations are also seen as an alternative to experiment. It is possible to simulate molecular systems that have not been synthesized yet, or phenomena inaccessible to experiments (huge temperature or pressure, time scales smaller than the femtosecond, evolutions on decades or more). Additionally, computations can serve for the laser control of molecular systems ([6]), and other emerging fields of high energy physics.

Other, apparently distant, fields also make an extensive use of molecular simulation. Rheology of complex fluids and more generally materials science were once focused at the macroscopic scale and based on purely macroscopic descriptions. They used to be far from molecular concerns. However, the accuracy needed in the quantitative evaluation of many properties (think e.g. of constitutive laws or slip boundary conditions) requires models to be more and more precise, involving the finest possible scales in the simulation. This eventually includes the molecular scale.

The last application field that we shall mention, besides the fields using the macroscopic impact of molecular simulation, regards the emerging field of nanotechnology. Nanosystems are indeed accessible today to a direct molecular simulation.

Overall, major technological challenges for the years to come may, or more properly stated must, be addressed by molecular simulation techniques. Examples are the detailed simulation of protein folding, and the description of the long time radiation damage of materials in nuclear power plants. To appreciate this ubiquity of molecular simulation, it is sufficient to consider the enormous proportion of computational time devoted to molecular simulation in the largest centers of computational resources worldwide.

**1.2. Relation to mathematics.** On the other hand, the interface of molecular simulation with mathematics is not yet comparable to the practical importance of the field.

Molecular simulation, and more precisely computational quantum chemistry, were born in the 1950s for molecular systems consisting of a few electrons. Contemporary methods and techniques now allow for the simulation of molecules of hundreds of electrons, modelled by very precise quantum models, up to samples of billions of particles modelled by molecular dynamics. This is an enormous success. The calculations are often surprisingly accurate, but also sometimes desperately inaccurate. Experts in chemistry have constantly improved the models and the methods. They have turned the field into an almighty tool. However, in many respects, molecular simulation is still an art. It relies upon a delicate mix of physical intuition, prag-

matic cleverness, and practical know-how. Mathematics has already provided with significant contributions to the theoretical understanding. Also, its companion fields, numerical analysis and scientific computing, have definitely improved the efficiency of the techniques. Yet, they all need to irrigate more molecular simulation. To state it otherwise, *there is an enormous gap between the sophistication of the models and the success of the numerical approaches used in practice and, on the other hand, the state of the art of their rigorous understanding.*

We are witnessing an evolution that is due to two different reasons.

First, the mathematical knowledge on the models is rather satisfactory. Efforts were initiated as early as the 1970s by pioneers such as E. H. Lieb, B. Simon, W. Thirring, Ch. Feffermann, focusing on fundamental theoretical issues. Questions were addressed about the well-posedness of the models, and the relation between the various models, in various asymptotic regimes. Researchers such as R. Benguria, J. P. Solovej, V. Bach, G. Friesecke, to only name a few, continued the effort over the years. Those were later joined by contributors following the impulsion given by P-L. Lions: M. J. Esteban, I. Catto, E. Séré, X. Blanc, M. Lewin, and the author. A number of researchers, experts in analysis, spectral theory, partial differential equations, evolution equations, now become involved in the field. The enclosed bibliography cites several of them.

Second, and as a natural follow-up to mathematical analysis, numerical analysis has indeed come into the picture. The numerical analysis of computational chemistry methods was a completely unexplored subject until the mid 1990s. Boosted by the state of the mathematical analysis, it is now a quickly developing topic. The work in this field was pioneered by E. Cancès. Researchers such as Y. Maday, M. Griebel, W. Hackbush, Ch. Lubich, W. E, well known for their contributions in various other fields of the engineering sciences, now get involved, along with their collaborators (G. Turinici, …) in electronic structure calculations or in molecular dynamics.

It is therefore a good time for assessing the field, and the issues and challenges ahead. Doing so might help to boost the research in the area.

The present contribution rapidly reviews some commonly used models and their mathematical nature, indicating the progress achieved over the last decades in their mathematical understanding. Questions of numerical analysis are also addressed. Important unsolved issues are emphasized. Owing to the evident space limitation, this review is not meant to be exhaustive: see [54], [55] for more comprehensive reviews, and [56] for a recent collection of various contributions. This is rather an invitation for mathematicians to get involved in the endeavour of placing the field on a firm mathematical ground.

## 2. Mathematical overview of the models of computational chemistry

**2.1. The Schrödinger equation.** For most applications of molecular simulation, the matter is described by an assembly of nuclei, which are point particles treated classically, equipped with electrons, which are light particles modelled by quantum mechanics. For systems of limited size, called *molecular systems*, there are $M$ nuclei, of charge $z_k$, located at $\bar{x}_k$, and $N$ electrons of unit charge. The finest models are called *ab initio* models since they only involve universal constants and no experimentally determined parameters. Assuming the molecular system nonrelativistic, placing it at zero temperature, and, for clarity of exposition, omitting the spin variable, the state of the electrons is modelled by the $N$-body Hamiltonian

$$H_e^{\bar{x}_1,\ldots,\bar{x}_M} = -\sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i} - \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{z_k}{|x_i - \bar{x}_k|} + \sum_{1\le i<j\le N} \frac{1}{|x_i - x_j|}, \tag{1}$$

where the terms respectively model the kinetic energy, the attraction between nuclei and electrons, the repulsion between electrons. Notice that the positions $\bar{x}_k$ of the nuclei are *parameters* of this operator. The *electronic ground-state* is by definition the minimizer of the energy:

$$W(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M) = \inf\left\{\langle\psi, H_e^{\bar{x}_1,\ldots,\bar{x}_M}\psi\rangle,\ \psi \in \mathcal{W}_N\right\} + \sum_{1\le k<l\le M} \frac{z_k\, z_l}{|\bar{x}_k - \bar{x}_l|}. \tag{2}$$

The variational space reads

$$\mathcal{W}_N = \left\{\psi \in \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3) : \int_{\mathbb{R}^{3N}} |\psi|^2 = 1,\ \int_{\mathbb{R}^{3N}} |\nabla\psi|^2 < +\infty\right\} \tag{3}$$

where the wedge product denotes the antisymmetrized tensor product (owing to the Pauli exclusion principle). The Euler–Lagrange equation of (2) is the (time-independent) Schrödinger equation

$$H_e^{\bar{x}_1,\ldots,\bar{x}_M}\psi = E\,\psi \tag{4}$$

where the energy $E$, lowest possible eigenvalue of $H_e^{\bar{x}_1,\ldots,\bar{x}_M}$ on $\mathcal{W}_N$ is called the ground-state energy. The resolution of (2) (or one approximation of it, which we will detail below) is at the core of any computational chemistry calculation, prior to any calculation related to excited states, energies, linear response, etc. We therefore focus on this problem here.

Analogously, a time-dependent version of the problem exists: then the *time-dependent* Schrödinger equation

$$i\frac{\partial}{\partial t}\psi = H_e^{\bar{x}_1,\ldots,\bar{x}_M}\psi \tag{5}$$

is to be solved. The treatment of the electronic problem is usually the *inner* loop of the simulation, the *outer* loop consisting of the treatment of the nuclei. In the static setting, this consists in solving the *molecular mechanics* problem (also termed *geometry optimization*): finding the configuration of nuclei that minimizes the overall energy, i.e. the minimizer of

$$\inf_{(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_M)\in\mathbb{R}^{3M}} W(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_M). \tag{6}$$

The time-dependent setting requires solving the equations of *molecular dynamics*, i.e. the Newton equations of motion for the nuclei:

$$m_k \frac{d^2}{dt^2}\bar{x}_k = -\nabla_{\bar{x}_k} W(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_M). \tag{7}$$

**2.2. Standard approximations.** Problem (2) is well explored mathematically: [37], [49]. In addition to their own interest, theoretical studies of (2) provide with useful practical information on the quantities (wavefunction and energy) to be evaluated in practice (see [48] and other works by the same authors). The practical bottleneck of quantum chemistry calculations is however that state-of-the-art numerical techniques only allow for (2) to be solved for ridiculously small numbers of electrons. Indeed, the dimension of the tensor product $\bigwedge_{i=1}^{N} L^2(\mathbb{R}^3)$ makes the problem untractable by usual techniques of scientific computing for the practically relevant numbers of electrons, say a few tens to thousands. The practice of computational chemistry is thus to *approximate* (2). The purpose of such approximations is to reduce the computational complexity of the problem, whilst providing the accuracy required by chemistry. The energy of molecular systems must indeed be determined within an incredibly demanding degree of accuracy (often termed the *chemical accuracy*). Energies such as (2) are typically $10^3$ to $10^6$ as large as the energy of an hydrogen bond. As the interest lies in the *difference* of energy between two systems, in order to determine which is the more stable one, the difficulty is challenging. Surprisingly, clever approximations do succeed in this task. We now review them. For more details on the analysis, implementation and efficiency of all the numerical techniques mentioned below, see [53].

In chemistry, approximations of (2) are schematically sorted into two categories.

*Wavefunctions methods* are used preferably by chemists, on small systems, when accuracy is the primary goal, and computational time is a secondary issue. The focus is on the *interaction* between electrons. The prototypical example is the Hartree–Fock model. The latter is the best known model in the mathematical community. The bottom line for deriving the Hartree–Fock model is a variational approximation of the set (3) by the subspace of wavefunctions $\psi$ that read as *determinants* (antisymmetrized products) of wavefunctions of *one* electron. More precisely, the Hartree–Fock problem reads

$$\inf\left\{E^{\mathrm{HF}}(\{\phi_i\}) : \phi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \phi_i\phi_j^* = \delta_{ij},\ 1\le i,j\le N\right\} \tag{8}$$

with

$$E^{\mathrm{HF}}(\{\phi_i\}) = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i(x)|^2\,dx + \int_{\mathbb{R}^3}\rho(x)\,V(x)\,dx + \frac{1}{2}\iint_{(\mathbb{R}^3)^2}\frac{\rho(x)\,\rho(y)}{|x-y|}\,dx\,dy - \frac{1}{2}\iint_{(\mathbb{R}^3)^2}\frac{|\tau(x,y)|^2}{|x-y|}\,dx\,dy, \tag{9}$$

with $V = -\sum_{k=1}^{M}\frac{z_k}{|\cdot-\bar{x}_k|}$, $\tau(x,y) = \sum_{i=1}^{N}\phi_i(x)\,\phi_i(y)^*$ and $\rho(x) = \tau(x,x) = \sum_{i=1}^{N}|\phi_i(x)|^2$, where the star denotes the complex conjugate. *Post–Hartree–Fock methods* consist in enlarging the variational space by considering *linear combinations* of determinants: *Configuration Interaction* (CI) methods, *Multiconfiguration Self Consistent Field* (MCSCF) methods. Nonvariational correction methods, mostly based on linear perturbation theory, are also employed: *Möller–Plesset*, *Coupled Cluster*.

On the other hand, *Density Functional Theory based methods* are used preferably for larger systems (and beyond for materials science), when computational time matters and wavefunctions methods are too expensive. They consist in rephrasing the problem (2) in terms of the electronic density

$$\rho(x) = N\int_{\mathbb{R}^{3(N-1)}}|\psi(x,x_2,x_3,\ldots,x_N)|^2\,dx_2\,dx_3\ldots dx_N.$$

Formally, a minimization problem of the type

$$\inf\left\{\mathcal{E}(\rho);\ \int_{\mathbb{R}^3}\rho(x)\,dx = N\right\} \tag{10}$$

is obtained. The idea has a rigorous theoretical grounding, but making it tractable in practice requires some approximation procedure. The energy $\mathcal{E}(\rho)$, which is a reformulation of $\langle\psi, H_e^{\bar{x}_1,\ldots,\bar{x}_M}\psi\rangle$, is not explicit. Adequately adjusting the parameters (and even the terms) of the approximate energy functional $\mathcal{E}(\rho)$ is an issue, sometimes controversial. Ancestors of DFT-based methods are Thomas–Fermi type theories, very well investigated mathematically (see [60], [63], [74] for reviews). The latter currently see a revival through *orbital-free* methods, which precisely consist in discretizing $\rho$ itself as the primary unknown. They therefore allow for the treatment of larger systems, notably for materials science applications.

The general trend is that DFT-based models are increasingly popular. A commonly used setting is the *Kohn–Sham Local Density Approximation* (KS-LDA) setting that explicitly reads as the minimization problem

$$\inf\left\{E^{\text{KS-LDA}}(\{\phi_i\}) : \phi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\phi_i(x)\phi_j^*(x)\,dx = \delta_{ij},\ 1\le i,j\le N\right\} \tag{11}$$

with

$$E^{\text{KS-LDA}}(\{\phi_i\}) = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i(x)|^2\,dx + \int_{\mathbb{R}^3}\rho(x)\,V(x)\,dx + \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho(x)\,\rho(y)}{|x-y|}\,dx\,dy - \int_{\mathbb{R}^3}F(\rho(x))\,dx, \tag{12}$$

where $F$ is a nonlinear function determined on chemical basis.

Reducing the complexity of (2) comes at a price: *nonlinearity*. Whereas the optimality equation (4) is a linear eigenvalue problem (in a high dimensional space), the equation to be solved for most of the approximations of (2) is a *nonlinear eigenvalue problem* (in a space of lower dimension, though). This is easily seen on the expressions (9) and (12). However they are derived and irrespective of their chemical meaning, the wavefunctions methods and DFT-based methods both lead to a *nonlinear eigenvalue problem*:

$$-\Delta\phi_j + \mathcal{W}(\phi_1,\ldots,\phi_N)\phi_j = \lambda_j\phi_j, \quad j = 1,\ldots,N, \tag{13}$$

where the $\lambda_j$ are the Lagrange multipliers of the constraints. Equations (13) are often called *Self-Consistent Field* (SCF) equations to emphasize the nonlinear feature, encoded in the operator $\mathcal{W}$.

There are many questions of mathematical interest. The existence of a minimizer (under appropriate physically relevant conditions) for several models related to HF and DFT-type approximations is now established. Very important contributions in this direction are [62], [61], [8], [65], [66], [73], [40], [59]. For most models of practical interest, the existence of a minimizer is known. In contrast, nothing is known on the uniqueness. A major reason for this is that almost all models of practical interest are *nonconvex*. The relation of these approximated models with the original model (2) has also been investigated, e.g. in [4], [38] for some physically relevant asymptotic regimes.

Mathematically, all problems arising in electronic structure theory are *nonlinear minimization problems* with possible lacks of compactness at infinity (most of them are posed on the whole space $\mathbb{R}^N$, and are subject to a constraint, see (8) and (11)). The Euler–Lagrange equation is a system of *nonlinear elliptic partial differential equations* such as (13). The ellipticity basically comes from the Laplacian operator, modelling the kinetic energy in (1). At one stage or another, *spectral theory* comes into the picture. More precisely, the spectral theory of *Schrödinger operators* $-\Delta + V$ often plays a key role. All this concerns the search for the ground state in the nonrelativistic setting, at zero temperature. When relativistic effects have to be accounted for, the Laplacian operator is replaced by the Dirac operator (unbounded from below), and the theoretical setting drastically changes. Important mathematical contributions on the relativistic setting are [36], [34], [35]. They have given birth to more efficient computational techniques. On the other hand, temperature effects may

also be accounted for, through the introduction of a statistics, see [64] for one of the rare mathematical studies. Like for temperature effects, the theory of excited states is not in a satisfactory state. Attempts to place the latter notion on a sound ground are [59], [23].

In the numerical practice, the problem is discretized using *Galerkin techniques*, and more precisely *spectral methods*. The basis functions used for discretization are typically gaussian approximations of the eigenfunctions of a hydrogen-like operator ($M = N = 1$ in (1)), or plane waves. The latter is very well adapted to solid state calculations. The former is incredibly efficient for calculations of molecular systems. A remarkable accuracy is reached with a limited number of basis functions. One reason why hydrogen-like basis functions outperform all other basis sets is that they are problem-dependent basis functions, which very well reproduce the exponential decrease at infinity and the cusp of wavefunctions at the point nuclei. More general purpose basis sets, such as finite elements, have difficulties in doing so, unless expensive mesh refinement techniques are employed. Finite-difference methods also exist, termed in this context *real-space* methods, but they are used for very specific applications, related to solid-state calculations.

After discretization, the equations are solved using *nonlinear optimization* techniques. Surprisingly, the problem is not addressed as a minimization problem, but in the form of the optimality equations (13). The latter is the only possible approach, considering the number of local minima, and despite the fact there is no theoretical basis for this. It reveals as an efficient approach, mostly because computations often benefit from prior calculations for adequately preconditioning the solution procedure. The algorithms in use for solving (13) are known as *SCF-algorithms*. Formally, they are elaborate variants of fixed-point iterations such as

$$-\Delta\phi_j^{n+1} + \mathcal{W}(\phi_1^n, \ldots, \phi_N^n)\phi_j^{n+1} = \lambda_j^{n+1}\phi_j^{n+1}, \quad j = 1, \ldots, N. \tag{14}$$

Their numerical analysis, initiated in [3], was performed only recently, see [18], [19] and [54] for a review. A rigorous mathematical insight into SCF-algorithms has led to definite improvements of their efficiency [52]. Alternative techniques may also be used. An original approach, based on *a posteriori error estimators* and related to Newton-type algorithms, is introduced in [69].

Notice that each inner loop of the nonlinear procedure involves a *linear* eigenvalue problem. This restricts the range of tractable systems (say typically that systems with a few hundreds of electrons can be standardly treated on a workstation). Ad hoc techniques may however be employed to broaden the spectrum of tractable systems. The latter are known as *linear scaling techniques*, for they significantly reduce the complexity of the diagonalization step, which in principle scales cubically with respect to the size of the system, see [12], [41], [42]. The bottom line for such a reduction is that the eigenelements are not explicitly needed: only the projector on the space spanned by the first $N$ eigenvectors is needed for the computation of all quantities of practical interest. The problem is thus rephrased so that an explicit diagonalization is avoided. Correspondingly, advanced techniques such as *Fast Multipole techniques* [51], are

used for assembling the huge matrices to be manipulated. Using a combination of such techniques, larger systems, consisting of thousands of electrons, may be treated on a workstation. The approach however still waits for a rigorous mathematical analysis.

The above description of the numerical approach concerns isolated molecular systems. Specific models are employed for the simulation of the liquid phase, and of the solid (crystalline) phase, respectively. In the former case, a commonly used setting is the *continuum model*: the molecule is placed in a cavity, surrounded by a dielectric medium modelling the solvent. Consequently, the Coulomb interaction potential appearing in $V$ (see (9) and (12)) is replaced by the Green function of electrostatics set on the cavity. *Integral equation methods* are utilized for the numerical resolution: [20], [21]. On the other hand, the modelling of the crystal phase corresponds to a periodic setting [2]: loosely speaking, the functions $\psi_j$ are indexed by a vector, i.e. for each $k$, $\psi_j^k(x)$ is the $j$-th eigenvector periodic in $x$ up to a phase factor $e^{-ik.x}$. In practice, the set of vectors $k$ is discretized, and the corresponding equations (13), now indexed by $k$, are solved. For the practical discretization and resolution of the equations, dedicated techniques are employed: see [53], [30]. Several theoretical issues regarding the rigorous derivation of the models for the crystalline phase have already been considered: see [62], [24], and other works by the same authors. Seminal contributions by L. Van Hove, F. Dyson, A. Lenard, D. Ruelle, E. Lieb, J. Lebowitz, B. Simon, Ch. Fefferman predated those. The bottom line is to justify the models of the solid phase proving they are the limits of models for molecular systems, as the system size grows. More generally, this is part of an enormous body of literature in mathematical physics addressing questions related to *thermodynamic limits*.

**2.3. Emerging approaches.** Wavefunctions methods and DFT-based methods are dominant computational methods. Apart from the main stream, there are three promising tracks followed either by chemists or mathematicians, that need to be advertised. They consist in addressing the problem (2) in its original form, without any approximation, in principle.

The first approach ([28]), actually almost as old as theoretical chemistry itself, is based on a rephrasing of the minimization problem in terms of the marginals

$$\gamma(x_1, x_2, x_1', x_2') = \int_{\mathbb{R}^{3(N-2)}} \psi(x_1, x_2, x_3, \ldots, x_N)\, \psi^*(x_1', x_2', x_3, \ldots, x_N)\, dx_3 \ldots dx_N,$$

called *second-order reduced density matrices*. This is possible because the operator (1) only involves the positions $x_i$ and $x_j$ of two electrons simultaneously.

The second approach (called *diffusion Monte-Carlo* in the specific context of chemistry) consists in determining the minimizer to problem (2) by solving the fictitious evolution equation

$$\frac{\partial \psi}{\partial t} + H_e \psi = 0,$$

using the Feynmann–Kac representation formula. Considering the long time limit provides with a strategy to evaluate (2), see [67], [22].

The third approach, advocated by some mathematicians ([44], [43]), consists in recognizing (4) as a high-dimensional partial differential equation and applying the techniques of *sparse tensor products*. The technique relies upon a theoretical framework set in this context in [78] (see also other works by the same author).

For all these three approaches, enormous theoretical and practical difficulties are still unsolved. For the first approach, the theoretical challenge is to determine the variational space for $\gamma$ corresponding to the variational space (3) for $\psi$. This is where *approximations* are again introduced. The story is not closed. Current techniques rely upon *semi-definite programming* or *Augmented Lagrangian methods* to solve the associated discretized problem. Somehow related to this, the difficulty for the last two approaches lies in the fermionic nature of the electrons: the wavefunction is constrained to be antisymmetric. In addition, the problem also requires appropriate techniques such as high-dimensional integration techniques, mainly based upon Monte-Carlo, or Quasi Monte-Carlo, techniques.

These three approaches are not in position today to compete with the other more classical ones, which have benefited over the past years from constant efforts shared by a huge community. They are however instances of approaches that may be turning points and may change the landscape of computational chemistry in the years to come.

All the above describes approaches to determine the electronic structure. As mentioned in the Introduction, this is most often the inner part of a calculation. The outer part concerns the nuclei, parameters of the inner calculation so far. In the static setting, the problem is usually to determine their optimal position, i.e. the most stable conformation. This is the molecular mechanics problem (6). In biology, such a problem is crucial. It is the well-known question of determining the 3-dimensional structure of the molecule (protein,…) under study. Techniques of *discrete optimization*, *combinatorial optimization*, in particular using *stochastics-based algorithms*, are employed. Notice that the mathematical question of the existence of such a most stable configuration is mostly open for all models of interest, in spite of outstanding contributions on academic models [25].

## 3. Dynamical problems and problems at larger scales

Regarding time-dependent problems, the evolution of the nuclei is again often considered classical. The Newton equations of motion (7) are solved. This is the extremely popular field of molecular dynamics. It is called *ab initio* when $W$ in the right-hand side of (7) is calculated *on-the-fly* from quantum mechanical models for the electronic structure (see [76] for a review), and *classical* when $W$ has a parameterized analytic form, fitted on previous calculations or experiments. Parameterized potentials reportedly work well in biological applications, but experience some difficulties for materials science applications.

For the explicit evaluation of $W$, a very common assumption in chemistry is *adiabaticity* (see [45], [46], [75] and other works by the same authors for mathematical discussions). When adiabaticity is assumed, $W$ ideally takes the form (2) and is computed using the static models and the techniques of Section 2.

In some cases such as collisions that involve electronic excited states with intersecting energy surfaces, the adiabatic approximation is not valid. Then equations (7) are coupled with the explicit time evolution of the electronic structure, simulated by (5), or one of its approximation (analogous to those of Section 2, adapted to the dynamics setting): see [17], [26], [50] for related mathematical studies.

A peculiarity to be borne in mind, which has a huge impact on the mathematical analysis, is that (7) is not only solved in order to determine the precise evolution of the system. Often, based on the *ergodicity* assumption, (7) serves as a tool for sampling the configuration space of the system in the microcanonical ensemble. Averages on this space are indeed related to quantities of macroscopic interest. Examples include the determination of the temperature, or the pressure, of a liquid system, or the determination of some mechanical properties such as the Young modulus of a crystalline solid.

Numerous challenging issues in numerical analysis arise from molecular dynamics. First, system (7) contains several, disproportionate, timescales. Think e.g. of bond lengths or angles oscillating either rapidly (i.e. at the femtosecond ($10^{-15}$ s) scale) or slowly (hundreds of femtoseconds). Adequate techniques must be employed: multi-timestep techniques, homogenization, damping of rapid degrees of freedom, integration of differential algebraic equations. Second, the integration of (7) over long times raises specific questions: geometric integration, backward error analysis, integration of Hamiltonian, symplectic, reversible systems, etc. For related questions, reference treatises or reviews in the numerical analysis literature are e.g. [11], [58], [47], [71, 13]. See also [72], [31] in the molecular dynamics community. Third, the longest timescales that may be reached using an explicit Hamiltonian dynamics are not sufficient to cover the practical needs. Say the limit is, in good cases, the microsecond and, more generally, the nanosecond. A major reason for this is that the evolution of the system basically consists of long period of oscillations around metastable sets (basins of energy), separated by rapid hoppings between these states (*simulation of infrequent events*). Techniques for reaching extremely long simulation times or for efficiently sampling the phase space are mandatory to complement standard molecular dynamics: *stochastic differential equations*, *Markov chains*, *path integrals*, etc. In addition, other ensembles than the microcanonical ensemble may be sampled by adequate deterministic modifications of Hamiltonian dynamics (*thermostated* equations of motion) or by stochastic equations (*Langevin dynamics*). See [27], [29], [33], [32], [77] and many other references by these authors and others, for examples of techniques. The above shows that molecular dynamics problems have a twofold *multiscale nature*: even on small time frames, they involve degrees of freedom with drastically different characteristic times, and in addition to this, the integration must be carried over extremely long times. This is a significant difficulty.

In spite of this, molecular dynamics simulation, along with acceleration techniques, is an extremely successful field and provides with impressively good quantitative results on some macroscopic quantities. Standard calculations on workstations simulate $10^8$ atoms over the nanosecond, record calculations largely outperform this. Here again, some practical and theoretical pitfalls remain and the mathematical understanding of the methods is to be improved. In a nutshell, one could say that it is not thoroughly understood *why molecular dynamics techniques perform so well, i.e. why averages calculated from erroneous or approximate trajectories are so close to the actual values of macroscopic quantities.*

The connection between the microscopic scale and the macroscopic scale is a broad subject. Calculations of ensemble averages using molecular dynamics and related techniques is one instance of it. Other questions concern the relation of molecular simulation with continuum mechanics. An example of a theoretical work in this direction is [9]. See also [10] for a review and references on the numerous practical applications, in particular applications related to computational materials science where strategies coupling molecular simulation techniques and continuum mechanics techniques are rapidly developping.

## 4. Trends

As briefly overviewed above, molecular simulation is an extremely rich application field of mathematics. Only a tiny part of the models and methods used in practice have been explored mathematically to date. There is much room for improvement in the mathematical understanding, the numerical analysis, the design of advanced techniques, to further enhance the field.

Some theoretical challenges concern the uniqueness of the ground state, the definition of excited states, the foundations of models at finite temperature, etc.

On the numerical side, current efforts in the mathematical community are directed towards the development of novel methods: sparse grids techniques, domain decomposition methods [7], stochastic methods for electronic structure calculations [22], methods for the determination of excited states [23], reduced basis methods [15], [16], parallel-in-time methods [5], stochastic methods for the computation of free energies [57], etc.

**Acknowledgements.** I wish to express my deepest gratitude to a number of colleagues for many years of collaboration on the various topics addressed in the present article. Thanks are due to P.-L. Lions, and X. Blanc, E. Cancès, I. Catto, M. J. Esteban, F. Legoll, T. Lelièvre, Y. Maday, G. Turinici.

## References


[1] Allen, M. P., Tildesley, D. J., *Computer simulation of liquids*. Oxford Science Publications, Oxford 1988.

[2] Ashcroft, N. W., Mermin, N. D., *Solid-State Physics*. Saunders College Publishing, New York 1976.

[3] Auchmuty, G., Jia, W., Convergent iterative methods for the Hartree eigenproblem. *RAIRO Modél. Math. Anal. Numér.* **28** (1994), 575–610.

[4] Bach, V., Error bound for the Hartree-Fock energy of atoms and molecules. *Comm. Math. Phys.* **147** (1992), 527–548.

[5] Baffico, L., Benard, S., Maday, Y., Turinici, G., Zérah, G., Parallel in time molecular dynamics simulations. *Phys. Rev. E.* **66** (2002), 057701.

[6] Bandrauk, A., Delfour, M., Le Bris, C. (eds.), *Quantum control: mathematical and numerical challenges*. CRM Proc. Lecture Notes 33, Amer. Math. Soc., Providence, RI, 2003.

[7] Barrault, M., Cancès, E., Hager, W. W., Le Bris, C., Multilevel domain decomposition for electronic structure calculations. *J. Comput. Phys.*, submitted

[8] Benguria, R., Brezis, H., Lieb, E. H., The Thomas-Fermi-von Weizsäcker theory of atoms and molecules. *Comm. Math. Phys.* **79** (1981), 167–180.

[9] Blanc, X., Le Bris, C., Lions, P.-L., From molecular models to continuum mechanics. *Arch. Ration. Mech. Anal.* **164** (2002), 341–381.

[10] Blanc, X., Le Bris, C., Lions, P.-L., Atomistic to Continuum limits for computational materials science. *Math. Model. Num. Anal.*, to appear.

[11] Borneman, F. A. *Homogenization in time of singularly perturbed mechanical systems*. Lectures Notes in Math. 1697, Springer-Verlag, Berlin 1998.

[12] Bowler, D. R., Miyazaki, T., Gillan, M. J., Recent progress in linear ab initio electronic structure techniques. *J. Phys. Condens. Matter* **14** (2002), 2781–2798.

[13] Budd, Ch., Piggott, M. D., Geometric Integration and its application. In *Handbook of Numerical Analysis* XI, North-Holland, Amsterdam 2004, 35–139.

[14] Cancès, E., Defranceschi, M., Kutzelnigg, W, Le Bris, C., Maday, Y., *Computational Quantum Chemistry: a Primer*. In *Handbook of Numerical Analysis X, Special volume, Computational Chemistry*, North-Holland, Amsterdam 2003, 3–270.

[15] Cancès, E., Le Bris, C., Maday, Y., Turinici, G., Towards reduced basis approaches in ab initio electronic computations. *J. Sci. Comput.* **17** (1-4) (2002), 461–469.

[16] Cancès, E., Le Bris, C., Maday, Y., Patera, T., Pau, G., Turinici, G., in preparation.

[17] Cancès, E., Le Bris, C., On the time-dependent electronic Hartree-Fock equations coupled with a classical nuclear dynamics. *Math. Models Methods Appl. Sci.* **9** (1999), 963–990.

[18] Cancès, E., Le Bris, C., On the convergence of SCF algorithms for the Hartree-Fock equations. *M2AN Math. Model. Numer. Anal.* **34** (2000), 749–774.

[19] Cancès, E., Le Bris, C., Can we outperform the DIIS approach for electronic structure calculations. *Int. J. Quantum Chem.* **79** (2000), 82–90.

[20] Cancès, E., Le Bris, C., Mennucci, B., Tomasi, J., Integral equation methods for molecular scale calculations in the liquid phase. *Math. Models Methods Appl. Sci.* **9** (1999) 35–44.

[21] Cancès, E., Mennucci, B., Tomasi, J., A new integral formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **101** (1997), 10506–10517.

[22] Cancès, E., Jourdain, B., Lelièvre, T., Quantum Monte-Carlo simulations of fermions: a mathematical analysis of the fixed-node approximation. *Math. Models Methods Appl. Sci.*, to appear.

[23] Cancès, E., Galicher, H., Lewin, M., Computing electronic structures: a new multiconfiguration approach for excited states. *J. Comp. Phys* **212** (2006), 73–98.

[24] Catto, I., Le Bris, C., Lions, P.-L., *Mathematical theory of thermodynamic limits: Thomas-Fermi type models*. Oxford Math. Monogr., Oxford University Press, New York 1998.

[25] Catto, I., Lions, P.-L., Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. I–IV. *Comm. Partial Differential Equations* **17** (1992), 1051–1110; **18** (1993), 305–354, 381–429, 1149–1159.

[26] Chadam, J. M., Glassey, R. T., Global existence of solutions to the Cauchy problem for time-dependent Hartree equations. *J. Math. Phys.* **16** (1975), 1122–1230.

[27] Chandler, D., Barrier crossings: classical theory of rare but important events. In *Computer Simulation of Rare Events and Dynamics of Classical and Quantum Condensed-Phase Systems – Classical and Quantum Dynamics in Condensed Phase Simulations* (Berne et al., eds.), World Scientific, Singapore 1998, 3–23.

[28] Coleman, A. J., Yukalov, V. I., *Reduced density matrices*. Lecture Notes in Chemistry 72, Springer-Verlag, Berlin 2000.

[29] Darve, E., Wilson, M., Pohorille, A., Calculating Free Energies Using Scaled-force Molecular Dynamics Algorithm. *Molecular Simulation* **28** (2002), 113.

[30] Defranceschi, M., Le Bris, C. (eds.), *Mathematical models and methods for ab initio quantum chemistry*. Lecture Notes in Chemistry 74, Springer-Verlag, Berlin 2001.

[31] Deuflhard, P., et al. (eds.), *Computational molecular dynamics: challenges, methods, ideas*. Lecture Notes in Comput. Sci. Eng. 4, Springer-Verlag, Berlin 1999.

[32] Deuflhard, P., Huisinga, W., Fischer, A., Schuette, Ch., Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **315** (1–3) (2000), 39–59.

[33] E, W., Ren, W., Vanden-Eijnden, E., String method for the study of rare events. *Phys. Rev. B* **66** (2002), 052301.

[34] Dolbeault, J., Esteban, M. J., Séré, E., On the eigenvalues of operators with gaps. Application to Dirac operators. *J. Funct. Anal.* **174** (2000), 208–226.

[35] Dolbeault, J., Esteban, M. J., Séré, E., A variational method for relativistic computations in atomic and molecular physics. *Int. J. Quant. Chem.* **93** (2003), 149–155.

[36] Esteban, M. J., Séré, E., Solutions of the Dirac-Fock equations for atoms and molecules. *Comm. Math. Phys.* **203** (1999), 499–530.

[37] Fefferman, Ch., The N-body problem in quantum mechanics *Commun. Pure Appl. Math.* **39** (1986), S67–S109.

[38] Fefferman, Ch., Seco, L. A., On the energy of a large atom. *Bull. Amer. Math. Soc.* **23** (1990), 525–530.

[39] Frenkel, D., Smit, B., *Understanding molecular simulation*. Second edition, Academic Press, San Diego, CA, 2002.

[40] Friesecke, G., The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions. *Arch. Ration. Mech. Anal.* **169** (2003), 35–71.

[41] Galli, G., Large scale electronic structure calculations using linear scaling methods. *Phys. Stat. Sol.* (b) **217** (2000), 231–249.

[42] Goedecker, S., Linear scaling electronic structure methods. *Rev. Mod. Phys.* **71** (1999), 1085–1123.

[43] Griebel, M., Hamaekers, J. Sparse grids for the Schrödinger equation, submitted.

[44] Hackbusch, W., The efficient computation of certain determinants arising in the treatment of Schrödinger's equations. *Computing* **67** (2001), 35–56.

[45] Hagedorn, G. A., Crossing the interface between Chemistry and Mathematics. *Notices Amer. Math. Soc.* **43** (3) (1996), 297–299.

[46] Hagedorn, G. A., Joye, A., Molecular propagation through small avoided crossings of electron energy levels. *Rev. Math. Phys.* **11** (1999), 41–101.

[47] Hairer, E., Lubich, Ch., Wanner, G., *Geometric numerical integration*. Springer Ser. Comput. Math. 31, Springer-Verlag, Berlin 2002.

[48] Hoffmann-Ostenhof, M., Hoffmann-Ostenhof, T., Sørensen, T. Electron wavefunctions and densities for atoms. *Ann. Henri Poincaré* **2** (2001), 77–100.

[49] Hunziker, W. and Sigal, I. M., The quantum $N$-body problem. *J. Math. Phys.* **41** (2000), 3348–3509.

[50] Koch, O. and Lubich, Ch., Regularity of the multi-configuration time-dependent hartree approximation in quantum molecular dynamics, *M2AN Math. Model. Numer. Anal.*, submitted.

[51] Kudin, K., Scuseria, G. E., A fast multipole algorithm for the efficient treatment of the Coulomb problem in electronic structure calculations of periodic systems with Gaussian orbitals. *Chem. Phys. Lett.* **289** (1998), 611–616.

[52] Kudin, K., Scuseria, G.E., Cancès, E., A black-box self-consistent field convergence algorithm: one step closer. *J. Chem. Phys.* **116** (2002), 8255–8261.

[53] Le Bris, C. (ed.), *Handbook of Numerical Analysis* X. Special volume: Computational Chemistry, North-Holland, Amsterdam 2003.

[54] Le Bris, C., Computational chemistry from the perspective of numerical analysis. *Acta Numer.* **14** (2005), 363–444.

[55] Le Bris, C., Lions, P. L., From atoms to crystals: a mathematical journey. *Bull. Amer. Math. Soc.* **42** (2005), 291–363.

[56] Le Bris, C., Patera, A. T. (eds.), Computational chemistry. A special issue of *M2AN Math. Model. Numer. Anal.*, to appear.

[57] Le Bris, C., Lelièvre, T., Vanden-Eijnden, E., Work in progress.

[58] Leimkuhler, B., Reich, S., *Simulating Hamiltonian dynamics*. Cambridge Monogr. Appl. Comput. Math. 14, Cambridge University Press, Cambridge 2004.

[59] Lewin, M., Solutions of the Multiconfiguration Equations in Quantum Chemistry. *Arch. Ration. Mech. Anal.* **171** (1) (2004), 83–114.

[60] Lieb, E. H., The stability of matter: from atoms to stars. *Bull. Amer. Math. Soc.* **22** (1990), 1–49.

[61] Lieb, E. H., Simon, B., The Hartree-Fock theory for Coulomb systems. *Comm. Math. Phys.* **53** (1977), 185–194.

[62] Lieb, E. H., Simon, B., The Thomas-Fermi theory of atoms, molecules and solids. *Adv. Math.* **23** (1977), 22–116.

[63] Lieb, E. H., Thomas-Fermi and related theories of atoms and molecules. *Rev. Mod. Phys.* **53** (1981), 603–641.

[64] Lions, P.-L., Hartree-Fock and related equations. In *Nonlinear partial differential equations and their applications*, Collège de France Seminar, Vol. IX (Paris, 1985–1986), Pitman Res. Notes Math. Ser. 181, Longman Sci. Tech., Harlow 1988, 304–333.

[65] Lions, P.-L., Solutions of Hartree-Fock equations for Coulomb systems. *Comm. Math. Phys.* **109** (1987), 33–97.

[66] Lions, P.-L., The concentration-compactness principle in the calculus of variations. The locally compact case. I, II. *Ann. Inst. Henri Poincaré Anal. Non Linéaire* **1** (1984), 109–145, 223–283.

[67] Lions, P.-L., Remarks on mathematical modelling in quantum chemistry. Computational Methods in Appl. Sci., Wiley, 1996, 22-23.

[68] McWeeny, R., *Methods of molecular quantum mechanics*. Second edition, Academic Press, 1992.

[69] Maday, Y., Turinici, G., Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations. *Numer. Math.* **94** (4) (2003), 739–770.

[70] Parr, R. G., Yang, W., *Density functional theory of atoms and molecules*. Oxford University Press, Oxford 1989.

[71] Sanz-Serna, J. M., Calvo, M. P., *Numerical Hamiltonian Problems*. Appl. Math. Math. Comput. 7, Chapman and Hall, London 1994.

[72] Schlick, T., et al. (eds.), Computational Molecular Biophysics. *J. Comput. Phys.* **151** (1999), 1–421.

[73] Solovej, J. P., Universality in the Thomas-Fermi-von Weizsäcker model of atoms and molecules. *Comm. Math. Phys.* **129** (1990), 561–598.

[74] Spruch, L., Pedagogic notes on Thomas-Fermi theory (and on some improvements): atoms, stars and the stability of bulk matter. *Rev. Mod. Phys.* **63** (1991), 151–209.

[75] Teufel, S., *Adiabatic perturbation theory in quantum dynamics*. Lecture Notes in Math. 1821, Springer-Verlag, Berlin 2003.

[76] Tuckerman, M. E., Ab initio molecular dynamics: Basic concepts, current trends and novel applications. *J. Phys. Condens. Matter* **14** (2002), R1297–R1355.

[77] Voter, A. F., Sørensen, M. R., Accelerating atomistic simulations of defect dynamics: Hyperdynamics, parallel replica dynamics and temperature accelerated dynamics. *Mat. Res. Soc. Symp. Proc.* **538** (1999), 427–439.

[78] Yserentant, H., On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. *Numer. Math* **98** (4) (2004), 731–759.

CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, 77455 Marne-La-Vallée, France
and
INRIA Rocquencourt, MICMAC project, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
E-mail: lebris@cermics.enpc.fr

# Evolutionary dynamics of cooperation

Martin A. Nowak

**Abstract.** Cooperation means a donor pays a cost, $c$, for a recipient to get a benefit $b$. In evolutionary biology, cost and benefit are measured in terms of fitness. While mutation and selection represent the main 'forces' of evolutionary dynamics, cooperation is a fundamental principle that is required for every level of biological organization. Individual cells rely on cooperation among their components. Multi-cellular organisms exist because of cooperation among their cells. Social insects are masters of cooperation. Most aspects of human society are based on mechanisms that promote cooperation. Whenever evolution constructs something entirely new (such as multi-cellularity or human language), cooperation is needed. Evolutionary construction is based on cooperation. I will present five basic principles for the evolution of cooperation, which arise in the theories of kin selection, direct reciprocity, indirect reciprocity, graph selection and group selection.

**Mathematics Subject Classification (2000).** 92B05.

**Keywords.** Mathematical biology, evolutionary dynamics, kin selection, evolutionary graph theory, indirect reciprocity, Prisoner's Dilemma.

## 1. Kin selection

In a pub conversation, J. B. S. Haldane, one of the founding fathers of a mathematical approach to biology, once remarked: 'I will jump into the river to save two brothers or eight cousins.' This insight was precisely formulated by William Hamilton many years later. He wrote a PhD thesis on this topic, submitted a long paper to the Journal of Theoretical Biology, disappeared into the Brazilian jungle and was world famous when he returned a few years later (Hamilton 1964a, b). The theory was termed 'kin selection' by John Maynard Smith (1964). The crucial equation is the following. Cooperation among relatives can be favored by natural selection if the coefficient of genetic relatedness, $r$, between the donor and the recipient exceeds the cost to benefit ratio of the altruistic act

$$r > c/b. \tag{1}$$

Kin selection theory has been tested in numerous experimental studies. Many cooperative acts among animals occur between close kin (Frank 1998, Hamilton 1998). The exact relationship between kin selection and other mechanisms such as group selection and spatial reciprocity, however, remains unclear. A recent study even suggests that much of cooperation in social insects is due to group selection rather than kin selection (Wilson & Hölldobler 2005).

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

## 2. Direct reciprocity

In 1971, Robert Trivers published a landmark paper entitled 'The evolution of reciprocal altruism' (Trivers 1971). Trivers analyzed the question of how natural selection could lead to cooperation between unrelated individuals. He discusses three biological examples: cleaning symbiosis in fish, warning calls in birds and human interactions. Trivers cites Luce & Raiffa (1957) and Rapoport & Chammah (1965) for the Prisoner's Dilemma, which is a game where two players have the option to cooperate or to defect. If both cooperate they receive the 'reward', $R$. If both defect they receive the 'punishment', $P$. If one cooperates and the other defects, then the cooperator receives the 'sucker's payoff', $S$, while the defector receives the 'temptation', $T$. The PD is defined by the ranking $T > R > P > S$.

Would you cooperate or defect? Assuming the other person will cooperate it is better to defect, because $T > R$. Assuming the other person will defect it is also better to defect, because $P > S$. Hence, no matter what the other person will do it is best to defect. If both players analyze the game in this 'rational' way then they will end up defecting. The dilemma is that they both could have received a higher payoff if they had chosen to cooperate. But cooperation is 'irrational'.

We can also imagine a population of cooperators and defectors and assume that the payoff for each player is determined by many random interactions with others. Let $x$ denote the frequency of cooperators and $1-x$ the frequency of defectors. The expected payoff for a cooperator is $f_C = Rx + S(1-x)$. The expected payoff for a defector is $f_D = Tx + P(1-x)$. Therefore, for any $x$, defectors have a higher payoff than cooperators. In evolutionary game theory, payoff is interpreted as fitness. Successful strategies reproduce faster and outcompete less successful ones. Reproduction can be cultural or genetic. In the non- repeated PD, in a well mixed population, defectors will outcompete cooperators. Natural selection favors defectors.

Cooperation becomes an option if the game is repeated. Suppose there are $m$ rounds. Let us compare two strategies, 'always defect' (ALLD), and GRIM, which cooperates on the first move, then cooperates as long as the opponent cooperates, but permanently switches to defection if the opponent defects once. The expected payoff for GRIM versus GRIM is $nR$. The expected payoff for ALLD versus GRIM is $T + (m-1)P$. If $nR > T + (m-1)P$ then ALLD cannot spread in a GRIM population when rare. This is an argument of evolutionary stability. Interestingly, Trivers (1971) quotes 'Hamilton (pers. commun.)' for this idea.

A small problem of the above analysis is that given a known number of rounds it is best to defect in the last round and by backwards induction it is also best to defect in the penultimate round and so on. Therefore, it is more natural to consider a repeated game with a probability $w$ of having another round. In this case, the expected number of rounds is $1/(1-w)$, and GRIM is stable against invasion by ALLD provided $w > (T-R)/(T-P)$.

We can also formulate the PD as follows. The cooperator helps at a cost, $c$, and the other individual receives a benefit $b$. Defectors do not help. Therefore we have

$T = b$, $R = b - c$, $P = 0$ and $S = -c$. The family of games that is described by the parameters $b$ and $c$ is a subset of all possible Prisoner's Dilemma games as long as $b > c$. For the repeated PD, we find that ALLD cannot invade GRIM if

$$w > c/b. \tag{2}$$

The probability of having another round must exceed the cost to benefit ratio of the altruistic act (Axelrod & Hamilton 1981, Axelrod 1984).

Thus, the repeated PD allows cooperation, but the question arises – what is a good strategy for playing this game? This question was posed by the political scientist, Robert Axelrod. In 1979, he decided to conduct a tournament of computer programs playing the repeated PD. He received 14 entries, from which the surprise winner was tit-for-tat (TFT), the simplest of all strategies that were submitted. TFT cooperates in the first move, and then does whatever the opponent did in the previous round. TFT cooperates if you cooperate, TFT defects if you defect. It was submitted by the game theorist Anatol Rapoport (who is also the co-author of the book Rapoport & Chammah,1965). Axelrod analyzed the events of the tournament, published a detailed account and invited people to submit strategies for a second championship. This time he received 63 entries. John Maynard Smith submitted 'tit-for-two-tats', a variant of TFT which defects only after the opponent has defected twice in a row. Only one person, Rapoport, submitted TFT, and it won again. At this time, TFT was considered to be the undisputed world champion in the heroic world of the repeated PD.

But one weakness became apparent very soon (Molander 1985, May 1987). TFT cannot correct mistakes. The tournaments were conducted without strategic noise. In the real world, 'trembling hands' and 'fuzzy minds' cause erroneous moves. If two TFT players interact with each other, a single mistake leads to a long sequence of alternating defection and cooperation. In the long run two TFT players get the same low payoff as two players who flip coins for every move in order to decide whether to cooperate or to defect. Errors destroy TFT.

In 1989, we began to conduct 'evolutionary tournaments' (Nowak & Sigmund 1992). Instead of inviting experts to submit programs, we asked mutation and selection to explore (some portion of) the strategy space of the repeated PD in the presence of noise. The initial random ensemble of strategies was quickly dominated by ALLD. If the opposition is nonsensical, it is best to defect. A large portion of the population began to adopt the ALLD strategy and everything seemed lost. But after some time, a small cluster of players adopted a strategy very close to TFT. If this cluster is sufficiently large, then it can increase in abundance, and the entire population swings from ALLD to TFT. Reciprocity (and therefore cooperation) has emerged. We can show that TFT is the best catalyst for the emergence of cooperation. But TFT's moment of glory was brief and fleeting. In all cases, TFT was rapidly replaced by another strategy. On close inspection, this strategy turned out to be 'generous-tit-for-tat' (GTFT) which always cooperates if the opponent has cooperated on the previous move, but sometimes (probabilistically) even cooperates when the opponent has defected. Natural selection had discovered 'forgiveness'.

After many generations, however, GTFT is undermined by unconditional cooperators, ALLC. In a society, where everybody is nice (using GTFT), there is almost no need to remember how to retaliate against a defection. A biological trait which is not used is likely to be lost by random drift. Birds that escape to islands without predators lose the ability to fly. Similarly, a GTFT population is softened and turns into an ALLC population.

Once most people play ALLC, there is an open invitation for ALLD to seize power. This is precisely what happens. The evolutionary dynamics run in cycles: from ALLD to TFT to GTFT to ALLC and back to ALLD. These oscillations of cooperative and defecting societies are a fundamental part of all our observations regarding the evolution of cooperation. Most models of cooperation show such oscillations. Cooperation is never a final state of evolutionary dynamics. Instead it is always lost to defection after some time and has to be re-established. These oscillations are also reminiscent of alternating episodes of war and peace in human history (Figure 1).

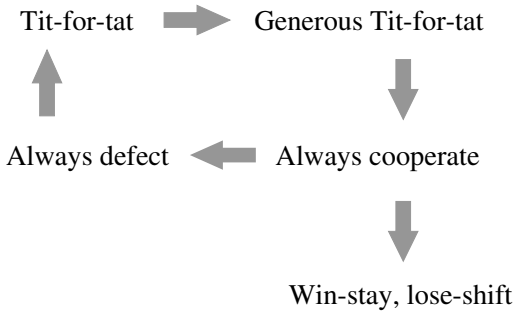



Figure 1. Evolutionary cycles of cooperation and defection. A small cluster of tit-for-tat (TFT) players or even a lineage starting from a single TFT player in a finite population can invade 'always defect' (ALLD). In fact, TFT is the most efficient catalyst for the first emergence of cooperation in an ALLD population. But in a world of 'fuzzy minds' and 'trembling hands', TFT is soon replaced by generous-tit-for-tat (GTFT), which can re-establish cooperation after occasional mistakes. If everybody uses GTFT, then 'always cooperate' (ALLC) is a neutral variant. Random drift leads to ALLC. An ALLC population invites invasion by ALLD. But ALLC is also dominated by 'win-stay, lose-shift' (WSLS), which leads to more stable cooperation than tit-for-tat-like strategies.

A subsequent set of simulations, exploring a larger strategy space, led to a surprise (Nowak & Sigmund 1993). The fundamental oscillations were interrupted by another strategy which seems to be able to hold its ground for a very long period of time. Most surprisingly, this strategy is based on the extremely simple principle of win-stay, lose-shift (WSLS). If my payoff is $R$ or $T$ then I will continue with the same move next round. If I have cooperated then I will cooperate again, if I have defected then I will defect again. If my payoff is only $S$ or $P$ then I will switch to the other move next round. If I have cooperated then I will defect, if I have defected then I will cooperate (Figure 2). If two WSLS strategists play each other, they cooperate

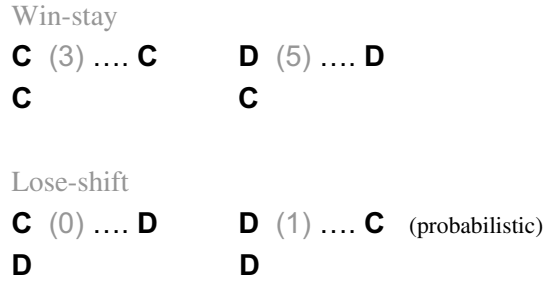



Figure 2. 'Win-stay, lose-shift' (WSLS) embodies a very simple principle. If you do well then continue with what you are doing. If you are not doing well, then try something else. Here we consider the Prisoner's Dilemma (PD) payoff values $R = 3$, $T = 5$, $P = 1$ and $S = 0$. If both players cooperate, you receive 3 points, and you continue to cooperate. If you defect against a cooperator, you receive 5 points, and you continue to defect. But if you cooperate with a defector, you receive 0 points, and therefore you will switch from cooperation to defection. If, on the other hand, you defect against a defector, you receive 1 point, and you will switch to cooperation. Your aspiration level is 3 points. If you get at least 3 points then you consider it a 'win' and you will 'stay' with your current choice. If you get less than 3 points, you consider it a 'loss' and you will 'shift' to another move. If $R > (T + P)/2$ (or $b/c > 2$) then WSLS is stable against invasion by ALLD. If this inequality does not hold, then our evolutionary simulations lead to a stochastic variant of WSLS, which cooperates after a DD move only with a certain probability. This stochastic variant of WSLS is then stable against invasion by ALLD.

most of the time. If a defection occurs accidentally, then in the next move both will defect. Hereafter both will cooperate again. WSLS is a simple deterministic machine to correct stochastic noise. While TFT cannot correct mistakes, both GTFT and WSLS can correct mistakes. But WSLS has an additional ace in its hand. When WSLS plays ALLC it will discover after some time that ALLC does not retaliate. After an accidental defection, WSLS will switch to permanent defection. Therefore, a population of WSLS players does not drift to ALLC. Cooperation based on WSLS is more stable than cooperation based on tit-for-tat-like strategies. The repeated PD is mostly known as a story of tit-for-tat, but win-stay, lose-shift is a superior strategy in an evolutionary scenario with errors, mutation and many generations (Fudenberg & Maskin 1990, Nowak & Sigmund 1993).

Incidentally, WSLS is stable against invasion by ALLD if $b/c > 2$. If instead $1 < b/c < 2$ then a stochastic variant of WSLS dominates the scene; this strategy cooperates after a mutual defection only with a certain probability. Of course, all strategies of direct reciprocity, such as TFT, GTFT or WSLS can only lead to the evolution of cooperation if the fundamental inequality (2) is fulfilled.

## 3. Indirect reciprocity

While direct reciprocity embodies the idea 'You scratch my back and I scratch yours', indirect reciprocity suggests 'You scratch my back and I scratch someone else's'. Why should this work? Presumably I will not get scratched if it becomes known that I scratch nobody. Indirect reciprocity, in this view, is based on reputation (Nowak & Sigmund 1998a, b, 2005). But why should you care about what I do to a third person?

The main reason why economists and social scientists are interested in indirect reciprocity is because one-shot interactions between anonymous partners in a global market become increasingly common and tend to replace the traditional long-lasting associations and long-term interactions between relatives, neighbors, or members of the same village. A substantial part of our life is spent in the 'company of strangers', and many transactions are no longer face-to-face. The growth of e-auctions and other forms of e-commerce is based, to a considerable degree, on reputation and trust. The potential to exploit such trust raises what economists call moral hazards. How effective is reputation, especially if information is only partial?

Evolutionary biologists, on the other hand, are interested in the emergence of human societies, which constitutes the last (up to now) of the major transitions in evolution. In contrast to other eusocial species, such as bees, ants or termites, humans display a high degree of cooperation between non-relatives (Fehr & Fischbacher 2003). A considerable part of human cooperation is based on moralistic emotions, such as anger directed towards cheaters or the 'warm inner glow' felt after performing an altruistic action. Intriguingly, humans not only feel strongly about interactions which involve them directly, they also judge actions between third parties as evidenced by the contents of gossip. There are numerous experimental studies of indirect reciprocity based on reputation (Wedekind & Milinski 2000, Milinski et al. 2002, Wedekind & Braithwaite 2002, Seinen & Schramm 2006).

A simple model of indirect reciprocity (Nowak & Sigmund 1998a, b) assumes that, within a well-mixed population, individuals meet randomly, one in the role of the potential donor, the other as the potential recipient. Each individual experiences several rounds of this interaction in each role, but never with the same partner twice. A player can follow either an unconditional strategy, such as always cooperate or always defect, or a conditional strategy, which discriminates among the potential recipients according to their past interactions. In a simple example, a discriminating donor helps a recipient if her score exceeds a certain threshold. A player's score is 0 at birth, increases whenever that player helps and decreases whenever the player withholds help. Individual-based simulations and direct calculations show that cooperation based on indirect reciprocity can evolve provided the probability, $p$, of knowing the social score of another person exceeds the cost-to-benefit ratio of the altruistic act,

$$p > c/b. \tag{3}$$

The role of genetic relatedness that is crucial for kin selection is replaced by social acquaintanceship. In a fluid population, where most interactions are anonymous and

people have no possibility of monitoring the social score of others, indirect reciprocity has no chance. But in a socially viscous population, where people know one another's reputation, cooperation by indirect reciprocity can thrive (Nowak & Sigmund 1998a).

In a world of binary moral judgments (Nowak & Sigmund 1998b, Leimar & Hammerstein 2001, Panchanathan & Boyd 2003, Fishman 2003, Brandt & Sigmund 2004, 2005), there are four ways of assessing donors in terms of 'first-order assessment': always consider them as good, always consider them as bad, consider them as good if they refuse to give, or consider them as good if they give. Only this last option makes sense. Second-order assessment also depends on the score of the receiver; for example, it can be deemed good to refuse help to a bad person. There are 16 second-order rules. Third-order assessment also depends on the score of the donor; for example, a good person refusing to help a bad person may remain good, but a bad person refusing to help a bad person remains bad. There are 256 third-order assessment rules. We display three of them in Figure 3. Using the Scoring assessment rule,

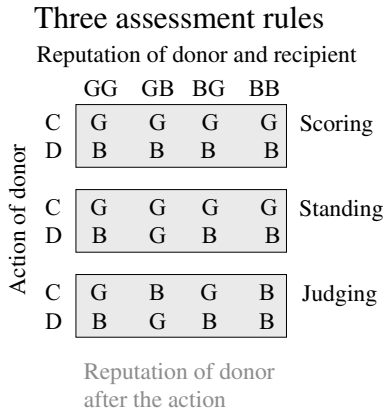



Figure 3. Assessment rules specify how an observer judges an interaction between a potential donor and a recipient. Here we show three examples of assessment rules in a world of binary reputation, good (G) and bad (B). For 'Scoring', cooperation (C) earns a good reputation and defection (D) earns a bad reputation. 'Standing' is very similar to Scoring, the only difference is that a 'good' donor can defect against a 'bad' recipient without losing his 'good' reputation. Note that Scoring is associated with costly punishment (Fehr & Gaechter 2002, Sigmund et al. 2001), whereas for Standing punishment of 'bad' recipients is cost-free. For 'Judging' it is 'bad' to help a 'bad' recipient.

cooperation, C, always leads to a good reputation, G, whereas defection, D, always leads to a bad reputation, B. Standing (Sugden 1986) is like Scoring, but it is not bad if a good donor defects against a bad recipient . With Judging, in addition, it is bad to cooperate with a bad recipient.

An action rule for indirect reciprocity prescribes giving or not giving, depending on the scores of both donor and recipient. For example, you may decide to help if the recipient's score is good or your own score is bad. Such an action might increase your

own score and therefore increase the chance of receiving help in the future. There are 16 action rules.

If we view a strategy as the combination of an action rule and an assessment rule, we obtain 4096 strategies. In a remarkable calculation, Ohtsuki & Iwasa (2004, 2005) analyzed all 4096 strategies and proved that only eight of them are evolutionarily stable under certain conditions and lead to cooperation (Figure 4). Both Standing

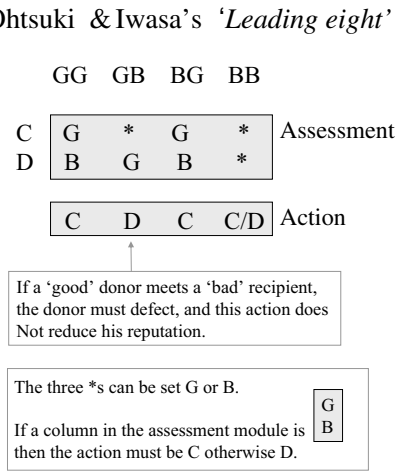



Figure 4. Ohtsuki & Iwasa (2004, 2005) analyzed the combination of $2^8 = 256$ assessment modules with $2^4 = 16$ action modules. This is a total of 4096 strategies. They found that 8 of these strategies can be evolutionarily stable and lead to cooperation, provided that everybody agrees on each other's reputation. (In general, uncertainty and incomplete information might lead to private lists of the reputation of others.) The three asterisks in the assessment module indicate a free choice between G and B. There are therefore $2^3 = 8$ different assessment rules which make up the 'leading eight'. The action module is built as follows: if the column in the assessment module is G and B, then the corresponding action is C, otherwise the action is D. Note that Standing and Judging are members of the leading eight, but neither Scoring nor Shunning is included.

and Judging belong to the leading eight, but not Scoring. We expect, however, that Scoring has a similar role in indirect reciprocity to that of tit-for-tat in direct reciprocity. Neither strategy is evolutionarily stable, but their ability to catalyze cooperation in adverse situations and their simplicity constitute their strength. In extended versions of indirect reciprocity in which donors can sometimes deceive others about the reputation of the recipient, Scoring is the 'foolproof' concept of 'I believe what I see'. Scoring judges the action and ignores the stories. There is also experimental evidence that in certain situations humans follow scoring rather than standing (Milinski et al. 2001).

In human evolution, there must have been a tendency to move from the simple cooperation promoted by kin or group selection to the strategic subtleties of direct and indirect reciprocity. Direct reciprocity requires precise recognition of individual

people, a memory of the various interactions one had with them in the past, and enough brain power to conduct multiple repeated games simultaneously. Indirect reciprocity, in addition, requires the individual to monitor interactions among other people, possibly judge the intentions that occur in such interactions, and keep up with the ever changing social network of the group. Reputation of players may not only be determined by their own actions, but also by their associations with others.

We expect that indirect reciprocity has coevolved with human language. On one hand, it is helpful to have names for other people and to receive information about how a person is perceived by others. On the other hand, a complex language is especially necessary if there are intricate social interactions. The possibilities for games of manipulation, deceit, cooperation and defection are limitless. It is likely that indirect reciprocity has provided the very selective scenario that led to cerebral expansion in human evolution.

## 4. Graph selection (or network reciprocity)

Game theory was invented by von Neumann and Morgenstern (1944) as a mathematical approach to understanding the strategic and economic decisions of humans. Hamilton (1967), Trivers (1971) and Maynard Smith & Price (1973) brought game theory to biology. Instead of analyzing the interaction between two rational players, evolutionary game theory explores the dynamics of a population of players under the influence of natural selection (Maynard Smith 1982). In the classical setting of the replicator equation, the population size is infinite and interactions are equally likely between any two individuals (Taylor & Jonker 1978, Hofbauer et al. 1979, Zeeman 1980). Each individual obtains an average payoff which is interpreted as biological fitness: strategies reproduce according to their payoff. Successful strategies spread and eliminate less successful ones. The payoff depends on the frequency of strategies in the population. Hence, natural selection is frequency dependent. The replicator equation is deeply connected to the concept of an evolutionarily stable strategy (ESS) or Nash equilibrium. In the framework of the replicator equation, an ESS cannot be invaded by any mutant strategy (Hofbauer & Sigmund 1998). For recent books on game theory and evolutionary game theory we refer to Fudenberg & Tirole 1991, Binmore 1994, Weibull 1995, Samuelson 1997, Fudenberg & Levine 1998, Hofbauer & Sigmund 1998, Gintis 2000, Cressman 2003. Recent reviews of evolutionary game dynamics are Hofbauer & Sigmund (2003) and Nowak & Sigmund (2004).

The traditional model of evolutionary game dynamics assumes that populations are well-mixed, which means that interactions between any two players are equally likely. More realistically, however, the interactions between individuals are governed by spatial effects or social networks. Let us therefore assume that the individuals of a population occupy the vertices of a graph (Nakamaru et al. 1997, 1998, Skyrms & Pemantle 2000, Abramson & Kuperman 2001, Ebel & Bornholdt 2002, Lieberman et al. 2005, Nakamaru & Iwasa 2005, Santos et al. 2005, Santos & Pacheco 2005).

The edges of the graph determine who interacts with whom (Figure 5). Consider a population of $N$ individuals consisting of cooperators and defectors. A cooperator

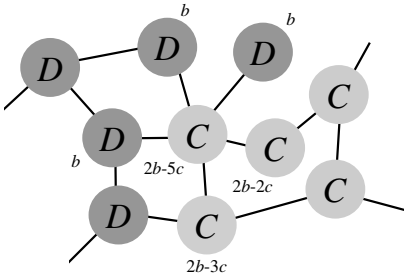



Figure 5. Games on graphs. The members of a population occupy the vertices of a graph (or social network). The edges denote who interacts with whom. Here we consider the specific example of cooperators, $C$, competing with defectors, $D$. A cooperator pays a cost, $c$, for every link. Each neighbor of a cooperator receives a benefit, $b$. The payoffs of some individuals are indicated in the figure. The fitness of each individual is a constant, denoting the 'baseline fitness', plus the payoff of the game. For evolutionary dynamics, we assume that in each round a random player is chosen to die, and the neighbors compete for the empty site in a mode proportional to their fitness. A simple rule emerges: if $b/c > k$ then selection favors cooperators over defectors. Here $k$ is the average number of neighbors per individual.

helps all individuals to whom it is connected. If a cooperator is connected to $k$ other individuals and $i$ of those are cooperators, then its payoff is $bi - ck$. A defector does not provide any help, and therefore has no costs, but it can receive the benefit from neighboring cooperators. If a defector is connected to $k$ other individuals and $j$ of those are cooperators, then its payoff is $bj$. Evolutionary dynamics are described by an extremely simple stochastic process: at each time step, a random individual adopts the strategy of one of its neighbors proportional to their fitness.

We note that stochastic evolutionary game dynamics in finite populations is sensitive to the intensity of selection. In general, the reproductive success (fitness) of an individual is given by a constant, denoting the baseline fitness, plus the payoff that arises from the game under consideration. Strong selection means that the payoff is large compared to the baseline fitness; weak selection means the payoff is small compared to the baseline fitness. It turns out that many interesting results can be proven for weak selection, which is an observation also well known in population genetics.

The traditional, well-mixed population of evolutionary game theory is represented by the complete graph, where all vertices are connected, which means that all individuals interact equally often. In this special situation, cooperators are always opposed by natural selection. This is the fundamental intuition of classical evolutionary game theory. But what happens on other graphs?

We need to calculate the probability, $\rho_C$, that a single cooperator, starting in a random position, turns the whole population from defectors into cooperators. If selection neither favors nor opposes cooperation, then this probability is $1/N$, which

is the fixation probability of a neutral mutant. If the fixation probability $\rho_C$ is greater than $1/N$, then selection favors the emergence of cooperation. Similarly, we can calculate the fixation probability of defectors, $\rho_D$.

A surprisingly simple rule determines whether selection on graphs favors cooperation. If

$$b/c > k, \tag{4}$$

then cooperators have a fixation probability greater than $1/N$ and defectors have a fixation probability less than $1/N$. Thus, for graph selection to favor cooperation, the benefit-to-cost ratio of the altruistic act must exceed the average degree, $k$, which is given by the average number of links per individual . This relationship can be shown with the method of pair-approximation for regular graphs, where all individuals have exactly the same number of neighbors (Ohtsuki et al. 2006). Regular graphs include cycles, all kinds of spatial lattices and random regular graphs. Moreover, computer simulations suggest that the rule $b/c > k$ also holds for non-regular graphs such as random graphs and scale free networks. The rule holds in the limit of weak selection and $k \ll N$. For the complete graph, $k = N$, we always have $\rho_D > 1/N > \rho_C$.

The basic idea is that natural selection on graphs (in structured populations) can favor unconditional cooperation without any need of strategic complexity, reputation or kin selection.

Games on graphs grew out of the earlier tradition of spatial evolutionary game theory (Nowak & May 1992, Herz 1994, Killingback & Doebeli 1996, Mitteldorf & Wilson 2000, Hauert et al. 2002, Le Galliard et al 2003, Hauert & Doebeli 2004, Szabo & Vukov 2004) and investigations of spatial models in ecology (Durrett & Levin 1994a, b, Hassell et al. 1994, Tilman & Kareiva 1997, Neuhauser 2001) and spatial models in population genetics (Wright 1931, Fisher & Ford 1950, Maruyama 1970, Slatkin 1981, Barton 1993, Pulliam 1988, Whitlock 2003).

## 5. Group selection

The enthusiastic approach of early group selectionists to explain the evolution of cooperation entirely from this one perspective (Wynne-Edwards 1962) has met with vigorous criticism (Williams 1966) and has led to a denial of group selection for decades. Only an embattled minority of scientists defended the approach (Eshel 1972, Wilson 1975, Matessi & Jayakar 1976, Wade 1976, Uyenoyama & Feldman 1980, Slatkin 1981, Leigh 1983, Szathmary & Demeter 1987). Nowadays, however, it seems clear that group selection acts as a powerful mechanism for the promotion of cooperation (Sober & Wilson 1998, Keller 1999, Michod 1999, Swenson et al. 2000, Kerr & Godfrey-Smith 2002, Paulsson 2002, Boyd & Richerson 2002, Bowles & Gintis 2004, Traulsen et al. 2005). We only have to make sure that its basic requirements are fulfilled in a particular situation (Maynard Smith 1976). We would like to illustrate exactly what these requirements are through the use of a simple model (Traulsen & Nowak 2006).

Imagine a population of individuals subdivided into groups. For simplicity, we assume the number of groups is constant and given by $m$. Each group contains between one and $n$ individuals. The total population size can fluctuate between the bounds $m$ and $nm$. Again, there are two types of individuals, cooperators and defectors. Individuals interact with others in their group and thereby receive a payoff. At each time step a random individual from the entire population is chosen proportional to payoff in order to reproduce. The offspring is added to the same group. If the group size is less than or equal to $n$ nothing else happens. If the group size, however, exceeds $n$ then with probability $q$ the group splits into two. In this case, a random group is eliminated (in order to maintain a constant number of groups). With probability $1-q$, the group does not divide, but instead a random individual from that group is eliminated (Figure 6). This minimalist model of multi-level selection has some interesting fea-

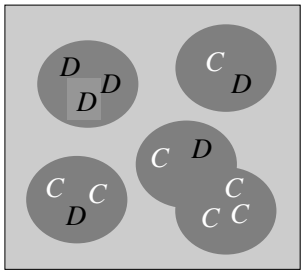



Figure 6. A simple model of group selection. A population consists of $m$ groups of maximum size $n$. Individuals interact with others in their group in the context of an evolutionary game. Here we consider the game between cooperators, $C$, and defectors, $D$. For reproduction, individuals are chosen from the entire population with a probability proportional to their payoff. The offspring is added to the same group. If a group reaches the maximum size, $n$, then it either splits in two or a random individual from that group is eliminated. If a group splits, then a random group dies, in order to keep the total population size constant. This meta-population structure leads to the emergence of two levels of selection, although only individuals reproduce.

tures. Note that the evolutionary dynamics are entirely driven by individual fitness. Only individuals are assigned payoff values. Only individuals reproduce. Groups can stay together or split (divide) when reaching a certain size. Groups that contain fitter individuals reach the critical size faster and therefore split more often. This concept leads to selection among groups, even though only individuals reproduce. The higher level selection emerges from lower level reproduction. Remarkably, the two levels of selection can oppose each other.

As before, we can compute the fixation probabilities, $\rho_C$ and $\rho_D$, of cooperators and defectors in order to check whether selection favors one or the other. If we add a single cooperator to a population of defectors, then this cooperator must first take over a group. Subsequently the group of cooperators must take over the entire population. The first step is opposed by selection, the second step is favored by selection. Hence,

we need to find out if the overall fixation probability is greater to or less than what we would obtain for a neutral mutant. An analytic calculation is possible in the limit $q \ll 1$ where individuals reproduce much more rapidly than groups divide. In this case, most of the groups are at their maximum size and hence the total population size is almost constant and given by $N = nm$. We find that selection favors cooperators and opposes defectors, $\rho_C > 1/N > \rho_D$, if

$$\frac{b}{c} > 1 + \frac{n}{m-2}. \tag{5a}$$

This result holds for weak selection. Smaller group sizes and larger numbers of competing groups favor cooperation. We also notice that the number of groups, $m$, must exceed two. There is an intuitive reason for this threshold. Consider the case of $m = 2$ groups with $n = 2$ individuals. In a mixed group, the cooperator has payoff $-c$ and the defector has payoff $b$. In a homogeneous group, two cooperators have payoff $b - c$, while two defectors have payoff 0. Thus the disadvantage for cooperators in mixed groups cannot be compensated for by the advantage they have in homogeneous groups. Interestingly, however, for larger splitting probabilities, $q$, we find that cooperators can be favored even for $m = 2$ groups. The reason is the following: for very small $q$, the initial cooperator must reach fixation in a mixed group; but for larger $q$, a homogeneous cooperator group can also emerge if a mixed group splits giving rise to a daughter group that has only cooperators. Thus, larger splitting probabilities make it easier for cooperation to emerge.

Let us also consider the effect of migration between groups. The average number of migrants accepted by a group during its life-time is denoted by $z$. We find that selection favors cooperation provided

$$\frac{b}{c} > 1 + z + \frac{n}{m}. \tag{5b}$$

In order, to derive this condition we have assumed weak selection and $q \ll 1$, as before, but also that both the numbers of groups, $m$, and the maximum group size, $n$, are much large than one.

Group selection (or multi-level selection) is a powerful mechanism for the evolution of cooperation if there are a large number of relatively small groups and migration between groups is not too frequent.

## 6. Conclusion

I have presented five simple (Equations 1–5) rules that determine whether particular mechanisms can promote the evolution of cooperation. In all five theories, $b$ is the benefit for the recipient and $c$ the cost for the donor of an altruistic act. The comparison of the five rules enables us to understand the crucial quantities that are responsible

for the natural selection of cooperation by the various mechanisms that have been proposed.

1. Kin selection leads to cooperation if $b/c > 1/r$, where $r$ is the coefficient of genetic relatedness between donor and recipient (Hamilton 1964a).

2. Direct reciprocity leads to cooperation if $b/c > 1/w$, where $w$ is the probability of playing another round in the repeated Prisoner's Dilemma (Axelrod & Hamilton 1981).

3. Indirect reciprocity leads to cooperation if $b/c > 1/q$, where $q$ is the probability to know the reputation of a recipient (Nowak & Sigmund 1998a).

4. Graph selection (or 'network reciprocity') leads to cooperation if $b/c > k$, where $k$ is the degree of the graph, that is the average number of neighbors (Ohtsuki et al. 2006).

5. Group selection leads to cooperation if $b/c > 1 + z + n/m$, where $z$ is the number of migrants accepted by a group during its life-time, $n$ is the group size and $m$ is the number of groups (Traulsen & Nowak 2006).

## References


[1] Abramson, G., and M. Kuperman. 2001. Social games in a social network. *Phys. Rev. E* **63** (3), 030901R.

[2] Axelrod, R. M. 1984. *The evolution of cooperation*. Basic Books, New York; reprint, Penguin, Harmondsworth 1989.

[3] Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. *Science* **211**, 1390–1396.

[4] Barton, N. 1993. The probability of fixation of a favoured allele in a subdivided population. *Genet. Res.* **62**, 149–158.

[5] Binmore, K. 1994. *Game theory and the social contract*, Vol. 1. *Playing fair*. MIT Press, Cambridge.

[6] Bowles, S., and H. Gintis. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoret. Population Biol.* **65**, 17–28.

[7] Boyd, R., and P. J. Richerson. 2002. Group beneficial norms can spread rapidly in a structured population. *J. Theoret. Biol.* **215**, 287–296.

[8] Brandt, H., and K. Sigmund. 2004. The logic of reprobation: Assessment and action rules for indirect reciprocity. *J. Theoret. Biol.* **231**, 475–486.

[9] Brandt, H., and K. Sigmund. 2005. Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl. Acad. Sci. USA* **102**, 2666–2670.

[10] Cressman, R. 2003. *Evolutionary dynamics and extensive form games*. MIT Press, Cambridge.

[11] Durrett, R., and S. A. Levin. 1994a. The importance of being discrete (and spatial). *Theoret. Population Biol.* **46**, 363–394.

[12] Durrett, R., and S. A. Levin. 1994b. Stochastic spatial models: A user's guide to ecological applications. *Philos. Trans. Roy. Soc. B* **343**, 329–350.

[13] Ebel, H., and S. Bornholdt. 2002. Coevolutionary games on networks. *Phys. Rev. E* **66** (5), 056118.

[14] Eshel, I. 1972. Neighbor effect and the evolution of altruistic traits. *Theoret. Population Biol.* **3**, 258–277.

[15] Fehr, E., and U. Fischbacher. 2003. The nature of human altruism. *Nature* **425**, 785–791.

[16] Fisher, R. A., and E. B. Ford. 1950. The Sewall Wright effect. *Heredity* **4**, 117–119.

[17] Fishman, M. A. 2003. Indirect reciprocity among imperfect individuals. *J. Theoret. Biol.* **225**, 285–292.

[18] Frank, S. A. 1998. *Foundations of social evolution*. Princeton University Press, Princeton, NJ.

[19] Fudenberg, D., and D. K. Levine. 1998. *The Theory of Learning in Games*. MIT Press, Cambridge.

[20] Fudenberg, D., and E. Maskin. 1990. Evolution and cooperation in noisy repeated games. *Amer. Econ. Rev.* **80**, 274–279.

[21] Fudenberg, D., and J. Tirole. 1991. *Game theory*. MIT Press, Cambridge.

[22] Gintis, H. 2000. *Game theory evolving*. Princeton University Press, Princeton, NJ.

[23] Hamilton, W. D. 1964a. The genetical evolution of social behaviour I. *J. Theoret. Biol.* **7**, 1–16.

[24] Hamilton, W. D. 1964b. The genetical evolution of social behaviour II. *J. Theoret. Biol.* **7**, 17–52.

[25] Hamilton, W. D. 1967. Extraordinary sex ratios. *Science* **156**, 477–488.

[26] Hamilton, W. D. 1998. *Narrow roads of gene land: The collected papers of W. D. Hamilton Volume 1: Evolution of social behaviour*. Oxford University Press, New York.

[27] Hassell, M. P., H. N. Comins, and R. M. May. 1994. Species coexistence and self-organizing spatial dynamics. *Nature* **370**, 290–292.

[28] Hauert, C., S. De Monte, J. Hofbauer, and K. Sigmund. 2002. Volunteering as red queen mechanism for cooperation in public goods games. *Science* **296**, 1129–1132.

[29] Hauert, C., and M. Doebeli. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* **428**, 643–646.

[30] Herz, A. V. M. 1994. Collective phenomena in spatially extended evolutionary games. *J. Theoret. Biol.* **169**, 65–87.

[31] Hofbauer, J., P. Schuster, and K. Sigmund. 1979. A note on evolutionarily stable strategies and game dynamics. *J. Theoret. Biol.* **81**, 609–612.

[32] Hofbauer, J., and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge University Press, Cambridge.

[33] Hofbauer, J., and K. Sigmund. 2003. Evolutionary game dynamics. *Bull. Amer. Math. Soc.* **40**, 479–519.

[34] Keller, L., ed. 1999. *Levels of selection in evolution*. Princeton University Press, Princeton, NJ.

[35] Kerr, B., and P. Godfrey-Smith. 2002. Individualist and multi-level perspectives on selection in structured populations. *Biol. Philos.* **17**, 477–517.

[36] Killingback, T., and M. Doebeli. 1996. Spatial evolutionary game theory: Hawks and Doves revisited. *Proc. Royal Soc. London B* **263**, 1135–1144.

[37] Le Galliard, J.-F., R. Ferrière, and U. Dieckmann. 2003. The adaptive dynamics of altruism in spatially heterogeneous populations. *Evolution* **57**, 1–17.

[38] Leigh, E. G. 1983. When does the good of the group override the advantage of the individual? *Proc. Natl. Acad. Sci. USA* **80**, 2985–2989.

[39] Leimar, O., and P. Hammerstein. 2001. Evolution of cooperation through indirect reciprocation. *Proc. Royal Soc. London B* **268**, 745–753.

[40] Lieberman, E., C. Hauert, and M. A. Nowak. 2005. Evolutionary dynamics on graphs. *Nature* **433**, 312–316.

[41] Luce, R. D., and H. Riaffa. 1957. *Games and Decisions*. John Wiley, New York, NY.

[42] Maruyama, T. 1970. Effective number of alleles in a subdivided population. *Theoret. Population Biol.* **1**, 273–306.

[43] Matessi, C., and S. D. Jayakar. 1976. Conditions for the evolution of altruism under Darwinian selection. *Theoret. Population Biol.* **9**, 360–387.

[44] May, R. M. 1987. More evolution of cooperation. *Nature* **327**, 15–17.

[45] Maynard Smith, J. 1964. Group selection and kin selection. *Nature* **63**, 20–29.

[46] Maynard Smith, J. 1976. Group selection. *Quart. Rev. Biol.* **201**, 145–147.

[47] Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge University Press, Cambridge.

[48] Maynard Smith, J., and G. R. Price. 1973. The logic of animal conflict. *Nature* **246**, 15–18.

[49] Michod, R. E. 1999. *Darwinian dynamics: Evolutionary transitions in fitness and individuality*. Princeton University Press, Princeton, NJ.

[50] Milinski, M., D. Semmann, T. C. M. Bakker, and H.-J. Krambeck. 2001. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. Royal Soc. London B* **268**, 2495–2501.

[51] Milinski, M., D. Semmann, and H.-J. Krambeck. 2002. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426.

[52] Mitteldorf, J., and D. S. Wilson. 2000. Population viscosity and the evolution of altruism. *J. Theoret. Biol.* **204**, 481–496.

[53] Molander, P. 1985. The optimal level of generosity in a selfish, uncertain environment. *J. Conflict Resolut.* **29**, 611–618.

[54] Nakamaru, M., and Y. Iwasa. 2005. The evolution of altruism by costly punishment in lattice-structured populations: Score-dependent viability versus score-dependent fertility. *Evol. Ecol. Res.*, in press.

[55] Nakamaru, M., H. Matsuda, and Y. Iwasa. 1997. The evolution of cooperation in a lattice-structured population. *J. Theoret. Biol.* **184**, 65–81.

[56] Nakamaru, M., H. Nogami, and Y. Iwasa. 1998. Score-dependent fertility model for the evolution of cooperation in a lattice. *J. Theoret. Biol.* **194**, 101–124.

[57] Neuhauser, C. 2001. Mathematical challenges in spatial ecology. *Notices Amer. Math. Soc.* **48**, 1304–1314.

[58] Nowak, M. A., and R. M. May. 1992. Evolutionary games and spatial chaos. *Nature* **359**, 826–829.

[59] Nowak, M. A., and K. Sigmund. 1992. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253.

[60] Nowak, M., and K. Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58.

[61] Nowak, M. A., and K. Sigmund. 1998a. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577.

[62] Nowak, M. A., and K. Sigmund. 1998b. The dynamics of indirect reciprocity. *J. Theoret. Biol.* **194**, 561–574.

[63] Nowak M. A., and K. Sigmund. 2004. Evolutionary dynamics of biological games. *Science* **303**, 793–799.

[64] Nowak, M. A., and K. Sigmund. 2005. Evolution of indirect reciprocity. *Nature*, in press.

[65] Ohtsuki, H., C. Hauert, E. Lieberman, and M. A. Nowak. 2006. A simple rule for the evolution of cooperation on graphs. *Nature*, in press.

[66] Ohtsuki, H., and Y. Iwasa. 2004. How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theoret. Biol.* **231**, 107–120.

[67] Ohtsuki, H., and Y. Iwasa. 2005. The leading eight: Social norms that can maintain cooperation by indirect reciprocity *J. Theoret. Biol.*, in press.

[68] Panchanathan, K., and R. Boyd. 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J. Theoret. Biol.* **224**, 115–126.

[69] Paulsson, J. 2002. Multileveled selection on plasmid replication. *Genetics* **161**, 1373–1384.

[70] Pulliam, H. R. 1988. Sources, sinks, and population regulation. *Amer. Nat.* **132**, 652–661.

[71] Rapoport, A., and A. M. Chammah. 1965. *Prisoner's dilemma*. University of Michigan Press, Ann Arbor, MI.

[72] Samuelson, L. 1997. *Evolutionary games and equilibrium selection*. MIT Press, Cambridge.

[73] Santos, F. C., and J. M. Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.* **95**, 098104.

[74] Santos, F. C., J. F. Rodrigues, and J. M. Pacheco. 2005. Graph topology plays a determinant role in the evolution of cooperation. *Proc. Royal Soc. London B* **273**, 51–55.

[75] Seinen, I., and A. Schram. 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.*, in press.

[76] Skyrms, B., and R. Pemantle. 2000. A dynamic model of social network formation. *Proc. Natl. Acad. Sci. USA* **97**, 9340–9346.

[77] Slatkin, M. 1981. Fixation probabilities and fixation times in a subdivided population. *Evolution* **35**, 477–488.

[78] Slatkin, M., and M. J. Wade. 1978. Group selection on a quantitative character. *Proc. Natl. Acad. Sci. USA* **75**, 3531–3534.

[79] Sober, E., and D. S. Wilson. 1998. *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press, Cambridge, MA.

[80] Sugden, R. 1986. *The economics of rights, co-operation and welfare*. Blackwell, Oxford.

[81] Swenson, W., D. S. Wilson, and R. Elias. 2000. Artificial ecosystem selection. *Proc. Natl. Acad. Sci. USA* **97**, 9110–9114.

[82] Szabó, G., and J. Vukov. 2004. Cooperation for volunteering and partially random partnerships. *Phys. Rev. E* **69** (3), 036107.

[83] Szathmáry, E., and L. Demeter. 1987. Group selection of early replicators and the origin of life. *J. Theoret. Biol.* **128**, 463–486.

[84] Takahashi, N., and R. Mashima. 2003. The emergence of indirect reciprocity: Is the standing strategy the answer? Center for the Study of cultural and ecological foundations of the mind, Hokkaido University, Japan, Working paper series No. 29.

[85] Taylor, P. D., and L. B. Jonker. 1978. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156.

[86] Tilman, D., and P. Karieva, eds. 1997. *Spatial ecology: The role of space in population dynamics and interspecific interactions*. Princeton University Press Monographs in Population Biology, Princeton University Press, Princeton, NJ.

[87] Traulsen, A., and M. A. Nowak. 2006. Emerging multi-level selection. Preprint.

[88] Traulsen, A., A. M. Sengupta, and M. A. Nowak. 2005. Stochastic evolutionary dynamics on two levels. *J. Theoret. Biol.* **235**, 393–401.

[89] Trivers, R. L. 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57.

[90] Uyenoyama, M., and M.W. Feldman. 1980. Theories of kin and group selection: A population genetics perspective. *Theoret. Population Biol.* **17**, 380–414.

[91] von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ.

[92] Wade, M. J. 1976. Group selection among laboratory populations of *Tribolium*. *Proc. Natl. Acad. Sci. USA* **73**, 4604–4607.

[93] Wedekind, C., and V. A. Braithwaite. 2002. The long term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12**, 1012–1015.

[94] Wedekind, C., and M. Milinski. 2000. Cooperation through image scoring in humans. *Science* **288**, 850–852.

[95] Weibull, J. 1995. *Evolutionary game theory*. MIT Press, Cambridge.

[96] Whitlock, M. 2003. Fixation probability and time in subdivided populations. *Genetics* **164**, 767–779.

[97] Williams, G. C. 1966. *Adaptation and natural selection*. Princeton University Press, Princeton, NJ.

[98] Wilson, E. O. 1975. *Sociobiology*. Harvard University Press, Cambridge, MA.

[99] Wilson, E. O., and B. Hölldobler. 1976. Eusociality: Origin and consequences. *Proc. Natl. Acad. Sci. USA* **102**, 13367–13371.

[100] Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**, 97–159.

[101] Wynn-Edwards, V. C. 1962. *Animal dispersion in relation to social behavior*. Oliver and Boyd, London.

[102] Zeeman, E. C. 1980. Population dynamics from game theory. In *Proceedings of an international conference on global theory of dynamical systems* (ed. by A. Nitecki and C. Robinson), Lecture Notes in Math. 819, Springer-Verlag, Berlin.

Program for Evolutionary Dynamics, Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, U.S.A.

E-mail: martin_nowak@harvard.edu

# Fractional Brownian motion: stochastic calculus and applications

David Nualart

**Abstract.** Fractional Brownian motion (fBm) is a centered self-similar Gaussian process with stationary increments, which depends on a parameter $H \in (0, 1)$ called the Hurst index. In this note we will survey some facts about the stochastic calculus with respect to fBm using a pathwise approach and the techniques of the Malliavin calculus. Some applications in turbulence and finance will be discussed.

**Mathematics Subject Classification (2000).** Primary 60H30; Secondary 60G18.

**Keywords.** Fractional Brownian motion, stochastic integrals, Malliavin calculus, Black–Scholes formula, stochastic volatility models.

## 1. Introduction

A real-valued stochastic process $X = \{X_t, t \geq 0\}$ is a family of random variables

$$X_t : \Omega \to \mathbb{R}$$

defined on a probability space $(\Omega, \mathcal{F}, P)$. The process $X$ is called *Gaussian* if for all $0 \leq t_1 < t_2 < \cdots < t_n$ the probability distribution of the random vector $(X_{t_1}, \ldots, X_{t_n})$ on $\mathbb{R}^n$ is normal or Gaussian. From the properties of the normal distribution it follows that the probability distribution of a Gaussian process is entirely determined by the mean function $\mathbb{E}(X_t)$ and the covariance function

$$\mathrm{Cov}(X_t, X_s) = \mathbb{E}((X_t - \mathbb{E}(X_t))(X_s - \mathbb{E}(X_s))),$$

where $\mathbb{E}$ denotes the mathematical expectation or integral with respect to the probability measure $P$.

One of the most important stochastic processes used in a variety of applications is the *Brownian motion* or *Wiener process* $W = \{W_t, t \geq 0\}$, which is a Gaussian process with zero mean and covariance function $\min(s, t)$. The process $W$ has independent increments and its formal derivative $\frac{dW_t}{dt}$ is used as input noise in dynamical systems, giving rise to stochastic differential equations. The stochastic calculus with respect to the Brownian motion, developed from the works of Itô in the forties, permits to formulate and solve stochastic differential equations.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

Motivated from some applications in hydrology, telecommunications, queueing theory and mathematical finance, there has been a recent interest in input noises without independent increments and possessing long-range dependence and self-similarity properties. Long-range dependence in a stationary time series occurs when the covariances tend to zero like a power function and so slowly that their sums diverge. The self-similarity property means invariance in distribution under a suitable change of scale. One of the simplest stochastic processes which is Gaussian, self-similar and it has stationary increments is fractional Brownian motion, which is a generalization of the classical Brownian motion. As we shall see later, the fractional Brownian motion possesses long-range dependence when its Hurst parameter is larger than 1/2.

In this note we survey some properties of the fractional Brownian motion, and describe different methods to construct a stochastic calculus with respect to this process. We will also discuss some applications in mathematical finance and in turbulence.

## 2. Fractional Brownian motion

A Gaussian process $B^H = \{B_t^H, t \geq 0\}$ is called *fractional Brownian motion* (fBm) of Hurst parameter $H \in (0, 1)$ if it has mean zero and the covariance function

$$\mathbb{E}(B_t^H B_s^H) = R_H(t, s) = \frac{1}{2}\big(s^{2H} + t^{2H} - |t - s|^{2H}\big). \tag{2.1}$$

This process was introduced by Kolmogorov [25] and studied by Mandelbrot and Van Ness in [30], where a stochastic integral representation in terms of a standard Brownian motion was established. The parameter $H$ is called Hurst index from the statistical analysis, developed by the climatologist Hurst [24], of the yearly water run-offs of Nile river.

The fractional Brownian motion has the following properties.

1. *Self-similarity:* For any constant $a > 0$, the processes $\{a^{-H} B_{at}^H, t \geq 0\}$ and $\{B_t^H, t \geq 0\}$ have the same probability distribution. This property is an immediate consequence of the fact that the covariance function (2.1) is homogeneous of order $2H$, and it can be considered as a "fractal property" in probability.

2. *Stationary increments:* From (2.1) it follows that the increment of the process in an interval $[s, t]$ has a normal distribution with zero mean and variance

$$\mathbb{E}\big((B_t^H - B_s^H)^2\big) = |t - s|^{2H}. \tag{2.2}$$

Hence, for any integer $k \geq 1$ we have

$$\mathbb{E}\big((B_t^H - B_s^H)^{2k}\big) = \frac{(2k)!}{k!2^k}|t - s|^{2Hk}. \tag{2.3}$$

Choosing $k$ such that $2Hk > 1$, Kolmogorov's continuity criterion and (2.3) imply that there exists a version of the fBm with continuous trajectories. Moreover, using Garsia–Rodemich–Rumsey lemma [19], we can deduce the following modulus of continuity for the trajectories of fBm: For all $\varepsilon > 0$ and $T > 0$, there exists a nonnegative random variable $G_{\varepsilon,T}$ such that $\mathbb{E}(|G_{\varepsilon,T}|^p) < \infty$ for all $p \geq 1$, and, almost surely,

$$\left|B_t^H - B_s^H\right| \leq G_{\varepsilon,T}|t-s|^{H-\varepsilon},$$

for all $s, t \in [0, T]$. In other words, the parameter $H$ controls the regularity of the trajectories, which are Hölder continuous of order $H - \varepsilon$, for any $\varepsilon > 0$.

For $H = 1/2$, the covariance can be written as $R_{1/2}(t, s) = \min(s, t)$, and the process $B^{1/2}$ is an ordinary Brownian motion. In this case the increments of the process in disjoint intervals are independent. However, for $H \neq 1/2$, the increments are not independent.

Set $X_n = B_n^H - B_{n-1}^H, n \geq 1$. Then $\{X_n, n \geq 1\}$ is a Gaussian stationary sequence with unit variance and covariance function

$$\begin{aligned}\rho_H(n) &= \frac{1}{2}\left((n+1)^{2H} + (n-1)^{2H} - 2n^{2H}\right)\\ &\approx H(2H-1)n^{2H-2} \to 0,\end{aligned}$$

as $n$ tends to infinity. Therefore, if $H > \frac{1}{2}$, $\rho_H(n) > 0$ for $n$ large enough and $\sum_{n=1}^{\infty} \rho_H(n) = \infty$. We say that the sequence $\{X_n, n \geq 1\}$ has *long-range dependence*. Moreover, this sequence presents an aggregation behavior which can be used to describe cluster phenomena. For $H < \frac{1}{2}$, $\rho_H(n) < 0$ for $n$ large enough and $\sum_{n=1}^{\infty} |\rho_H(n)| < \infty$. In this case, $\{X_n, n \geq 1\}$ can be used to model sequences with intermittency.

**2.1. Construction of the fBm.** In order to show the existence of the fBm we should check that the symmetric function $R_H(t, s)$ defined in (2.1) is nonnegative definite, that is,

$$\sum_{i,j=1}^{n} a_i a_j R_H(t_i, t_j) \geq 0 \tag{2.4}$$

for any sequence of real numbers $a_i$, $i = 1, \ldots, n$ and for any sequence $t_i \geq 0$. Property (2.4) follows from the integral representation

$$B_t^H = \frac{1}{C_1(H)} \int_{\mathbb{R}} \left[((t-s)^+)^{H-\frac{1}{2}} - ((-s)^+)^{H-\frac{1}{2}}\right] dW_s, \tag{2.5}$$

where $\{W(A), A \text{ Borel subset of } \mathbb{R}\}$ is a Brownian measure on $\mathbb{R}$ and

$$C_1(H) = \left(\int_0^{\infty} \left((1+s)^{H-\frac{1}{2}} - s^{H-\frac{1}{2}}\right)^2 ds + \frac{1}{2H}\right)^{\frac{1}{2}},$$

obtained by Mandelbrot and Van Ness in [30]. The stochastic integral (2.5) is well defined, because the function $f_t(s) = ((t-s)^+)^{H-\frac{1}{2}} - ((-s)^+)^{H-\frac{1}{2}}$, $s \in \mathbb{R}$, $t \geq 0$ satisfies $\int_{\mathbb{R}} f_t(s)^2 ds < \infty$. On the other hand, the right-hand side of (2.5) defines a zero mean Gaussian process such that

$$\mathbb{E}\big((B_t^H)^2\big) = t^{2H}$$

and

$$\mathbb{E}\big((B_t^H - B_s^H)^2\big) = (t-s)^{2H},$$

which implies that $B^H$ is an fBm with Hurst parameter $H$.

**2.2. *p*-variation of the fBm.** Suppose that $X = \{X_t, t \geq 0\}$ is a stochastic process with continuous trajectories. Fix $p > 0$. We define the $p$-variation of $X$ on an interval $[0, T]$ as the following limit in probability:

$$\lim_{n\to\infty} \sum_{j=1}^{n} \big| X_{\frac{jT}{n}} - X_{\frac{(j-1)T}{n}} \big|^p.$$

If the $p$-variation exists and it is nonzero a.s., then for any $q > p$ the $q$-variation is zero and for any $q < p$ the $q$-variation is infinite. For example, the 2-variation (or quadratic variation) of the Brownian motion is equal to the length of the interval $T$.

Rogers has proved in [40] that the fBm $B^H$ has finite $1/H$-variation equals to $c_p T$, where $c_p = \mathbb{E}(|B_1^H|^p)$. In fact, the self-similarity property implies that the sequence

$$\sum_{j=1}^{n} \big| B^H_{\frac{jT}{n}} - B^H_{\frac{(j-1)T}{n}} \big|^{1/H}$$

has the same distribution as

$$\frac{T}{n} \sum_{j=1}^{n} \big| B_j^H - B_{j-1}^H \big|^{1/H},$$

and by the Ergodic Theorem this converges in $L^1(\Omega)$ and almost surely to $\mathbb{E}(|B_1^H|^p)T$.

As a consequence, the fBm with Hurst parameter $H \neq 1/2$ is not a semimartingale. Semimartingales are the natural class of processes for which a stochastic calculus can be developed, and they can be expressed as the sum of a bounded variation process and a local martingale which has finite quadratic variation. The fBm cannot be a semimartingale except in the case $H = 1/2$ because if $H < 1/2$, the quadratic variation is infinite, and if $H > 1/2$ the quadratic variation is zero and the 1-variation is infinite.

Let us mention the following surprising result proved by Cheridito in [8]. Suppose that $\{B_t^H, t \geq 0\}$ is an fBm with Hurst parameter $H \in (0, 1)$, and $\{W_t, t \geq 0\}$ is an ordinary Brownian motion. Assume they are independent and set

$$M_t = B_t^H + W_t.$$

Then $\{M_t, t \geq 0\}$ is not a semimartingale if $H \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, \frac{3}{4}\right]$, and it is a semimartingale, equivalent in law to a Brownian motion on any finite time interval $[0, T]$, if $H \in \left(\frac{3}{4}, 1\right)$.

The $1/H$-variation of Wick stochastic integrals with respect to the fractional Brownian motion with parameter $H > 1/2$ has been computed by Guerra and Nualart in [20].

## 3. Stochastic calculus with respect to the fBm

The aim of the stochastic calculus is to define stochastic integrals of the form

$$\int_0^T u_t \, dB_t^H, \tag{3.1}$$

where $u = \{u_t, t \in [0, T]\}$ is some stochastic process. If $u$ is a deterministic function there is a general procedure to define the stochastic integral of $u$ with respect to a Gaussian process using the convergence in $L^2(\Omega)$. We will first review this general approach in the particular case of the fBm.

**3.1. Integration of deterministic processes.** Consider an fBm $B^H = \{B_t^H, t \geq 0\}$ with Hurst parameter $H \in (0, 1)$. Fix a time interval $[0, T]$ and denote by $\mathcal{E}$ the set of step functions on $[0, T]$. The integral of a step function of the form

$$\varphi_t = \sum_{j=1}^m a_j \mathbf{1}_{(t_{j-1}, t_j]}(t)$$

is defined in a natural way by

$$\int_0^T \varphi_t \, dB_t^H = \sum_{j=1}^m a_j (B_{t_j}^H - B_{t_{j-1}}^H).$$

We would like to extend this integral to a more general class of functions, using the convergence in $L^2(\Omega)$. To do this we introduce the Hilbert space $\mathcal{H}$ defined as the closure of $\mathcal{E}$ with respect to the scalar product

$$\langle \mathbf{1}_{[0,t]}, \mathbf{1}_{[0,s]} \rangle_{\mathcal{H}} = R_H(t, s).$$

Then the mapping $\varphi \longrightarrow \int_0^T \varphi_t \, dB_t^H$ can be extended to a linear isometry between $\mathcal{H}$ and the Gaussian subspace $H_T(B^H)$ of $L^2(\Omega, \mathcal{F}, P)$ spanned by the random variables $\{B_t^H, t \in [0, T]\}$. We will denote this isometry by $\varphi \longrightarrow B^H(\varphi)$.

We would like to interpret $B^H(\varphi)$ as the stochastic integral of $\varphi \in H_T(B^H)$ with respect to $B^H$ and to write $B^H(\varphi) = \int_0^T \varphi_t dB_t^H$. However, we do not know whether the elements of $\mathcal{H}$ can be considered as real-valued functions. This turns out to be true for $H < \frac{1}{2}$, but is false when $H > \frac{1}{2}$ (see Pipiras and Taqqu [38], [39]). We state without proof the following results about the space $\mathcal{H}$.

**3.1.1. Case $H > \frac{1}{2}$.** In this case the second partial derivative of the covariance function

$$\frac{\partial^2 R_H}{\partial t \partial s} = \alpha_H |t - s|^{2H-2},$$

where $\alpha_H = H(2H - 1)$, is integrable, and we can write

$$R_H(t, s) = \alpha_H \int_0^t \int_0^s |r - u|^{2H-2} \, du dr. \tag{3.2}$$

Formula (3.2) implies that the scalar product in the Hilbert space $\mathcal{H}$ can be written as

$$\langle \varphi, \psi \rangle_{\mathcal{H}} = \alpha_H \int_0^T \int_0^T |r - u|^{2H-2} \varphi_r \psi_u \, du dr \tag{3.3}$$

for any pair of step functions $\varphi$ and $\psi$ in $\mathcal{E}$.

As a consequence, we can exhibit a linear space of functions contained in $\mathcal{H}$ in the following way. Let $|\mathcal{H}|$ be the Banach space of measurable functions $\varphi \colon [0, T] \to \mathbb{R}$ such that

$$\|\varphi\|^2_{|\mathcal{H}|} = \alpha_H \int_0^T \int_0^T |r - u|^{2H-2} |\varphi_r| |\varphi_u| \, du dr < \infty.$$

It has been shown in [39] that the space $|\mathcal{H}|$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is not complete and it is isometric to a subspace of $\mathcal{H}$. The following estimate has been proved in [31] using Hölder and Hardy–Littlewood inequalities.

**Lemma 3.1.** *Let $H > \frac{1}{2}$ and $\varphi \in L^{\frac{1}{H}}([0, T])$. Then*

$$\|\varphi\|_{|\mathcal{H}|} \leq b_H \|\varphi\|_{L^{\frac{1}{H}}([0,T])}, \tag{3.4}$$

*for some constant $b_H$.*

Thus we have the embeddings

$$L^2([0, T]) \subset L^{\frac{1}{H}}([0, T]) \subset |\mathcal{H}| \subset \mathcal{H},$$

and Wiener-type integral $\int_0^T \varphi_t dB_t$ can be defined for functions $\varphi$ in the Banach space $|\mathcal{H}|$. Notice that we can integrate more functions that in the case of the Brownian motion, and the isometry property of the Itô stochastic integral is replaced here by the formula

$$\mathbb{E}\left(\left(\int_0^T \varphi_t \, dB_t^H\right)^2\right) = \alpha_H \int_0^T \int_0^T |r-u|^{2H-2} \varphi_r \varphi_u \, du dr = \|\varphi\|_{\mathcal{H}}^2.$$

**3.1.2. Case $H < \frac{1}{2}$.** In this case, one can show that $\mathcal{H} = I_{T-}^{\frac{1}{2}-H}(L^2([0,T]))$ (see [14] and Proposition 6 of [2]), where $I_{T-}^{\frac{1}{2}-H}$ is the right-sided fractional integral operator defined by

$$I_{T-}^{H-\frac{1}{2}}\varphi(t) = \frac{1}{\Gamma\left(H-\frac{1}{2}\right)} \int_t^T (s-t)^{H-\frac{3}{2}} \varphi_s \, ds.$$

This means that $\mathcal{H}$ is a space of functions. Moreover the norm of the Hilbert space $\mathcal{H}$ can be computed as follows:

$$\|\varphi\|_{\mathcal{H}}^2 = c_H^2 \int_0^T s^{1-2H} \left(D_{T-}^{\frac{1}{2}-H}(u^{H-\frac{1}{2}}\varphi_u)\right)^2(s) \, ds, \tag{3.5}$$

where $c_H$ is a constant depending on $H$ and $D_{T-}^{\frac{1}{2}-H}$ is the right-sided fractional derivative operator. The operator $D_{T-}^{\frac{1}{2}-H}$ is the inverse of $I_{T-}^{H-\frac{1}{2}}$, and it has the following integral expression:

$$D_{T-}^{\frac{1}{2}-H}\varphi(t) = \frac{1}{\Gamma\left(H+\frac{1}{2}\right)} \left( \frac{\varphi_t}{(T-t)^{\frac{1}{2}-H}} + \left(\frac{1}{2}-H\right) \int_t^T \frac{\varphi_t - \varphi_s}{(s-t)^{\frac{3}{2}-H}} \, ds \right). \tag{3.6}$$

The following embeddings hold:

$$C^\gamma([0,T]) \subset \mathcal{H} \subset L^{1/H}([0,T])$$

for any $\gamma > H - \frac{1}{2}$. The first inclusion is a direct consequence of formula (3.6), and the second one follows from Hardy–Littlewood inequality. Roughly speaking, in this case the fractional Brownian motion is more irregular than the classical Brownian motion, and some Hölder continuity is required for a function to be integrable. Moreover the computation of the variance of an integral using formula (3.5) is more involved.

**3.2. Integration of random processes.** Different approaches have been used in the literature in order to define stochastic integrals with respect to the fBm. Lin [26] and Dai and Heyde [13] have defined a stochastic integral $\int_0^T u_t dB_t^H$ as limit in $L^2$ of Riemann sums in the case $H > \frac{1}{2}$. The techniques of Malliavin calculus have been used to develop the stochastic calculus for the fBm starting from the pioneering

work of Decreusefond and Üstünel [14]. We refer to the works of Carmona and Coutin in [7], Alòs, Mazet and Nualart [1], [2], Alòs and Nualart [3], and the recent monograph by Hu [21], among others. We will first describe a path-wise approach based on Young integrals.

**3.2.1. Path-wise approach.** We can define $\int_0^T u_t\, dB_t^H$ using path-wise Riemann–Stieltjes integrals taking into account the results of Young in [43]. In fact, Young proved that the Riemann–Stieltjes integral $\int_0^T f_t\, dg_t$ exists, provided that $f, g\colon [0,T] \to \mathbb{R}$ are Hölder continuous functions of orders $\alpha$ and $\beta$ with $\alpha + \beta > 1$. Therefore, if $u = \{u_t, t \in [0,T]\}$ is a stochastic process with $\gamma$-Hölder continuous trajectories, where $\gamma > 1 - H$, then the Riemann–Stieltjes integral $\int_0^T u_t\, dB_t^H$ exists path-wise. That is for any element $\omega \in \Omega$, the integral $\int_0^T u_t(\omega)\, dB_t^H(\omega)$ exists as the point-wise limit of Riemann sums. In particular, if $H > 1/2$, the path-wise Riemann–Stieltjes integral $\int_0^T F(B_t^H)\, dB_t^H$ exists if $F$ is a continuously differentiable function. Moreover the following change of variables formula holds:

$$\Phi(B_t^H) = \Phi(0) + \int_0^t F(B_s^H)\, dB_s^H \tag{3.7}$$

if $\Phi' = F$.

In the case $\frac{1}{4} < H < \frac{1}{2}$, there is a path-wise approach to the stochastic integrals of the form

$$\int_0^T F(B_t^H)\, dB_t^H$$

using the theory of *rough paths analysis* introduced by Lyons in [27] (see also [28]). This theory has allowed Coutin and Qian [12] to show the existence of a solution and to prove the convergence of the Wong–Zakai approximations for stochastic differential equations driven by an fBm with Hurst parameter $H \in \left(\frac{1}{4}, \frac{1}{2}\right)$.

Nevertheless, unlike the case of the Itô stochastic integral with respect to the Brownian motion, the path-wise integral $\int_0^T F(B_t^H)\, dB_t^H$ does not have zero mean and there is no easy formula for its variance. We are going to explain how the techniques of Malliavin calculus allow us to compute the mean and the variance of this integral.

**3.3. Malliavin calculus for the fBm.** Let $B^H = \{B_t^H, t \geq 0\}$ be an fBm with Hurst parameter $H \in (0,1)$. The process $B^H$ is Gaussian and we can develop the corresponding stochastic calculus of variations or Malliavin calculus. The Malliavin calculus is an infinite dimensional differential calculus introduced by Malliavin in [29] to provide a probabilistic proof of Hörmander hypoellipticity theorem. The basic operators of Malliavin calculus are the derivative operator $D$ and its adjoint the divergence operator $\delta$. We refer to Nualart [32] and [33] for a detailed account of the Malliavin calculus and its application in the framework of the fBm.

Fix a time interval $[0, T]$. Let $\mathcal{S}$ be the set of elementary random variables of the form

$$F = f(B^H(\varphi_1), \dots, B^H(\varphi_n)), \tag{3.8}$$

where $n \geq 1$, $f \in C_p^\infty(\mathbb{R}^n)$ ($f$ and all its partial derivatives have polynomial growth order), and $\varphi_i \in \mathcal{H}$.

The *derivative operator* $D$ of an elementary random variable $F$ of the form (3.8) is defined as the $\mathcal{H}$-valued random variable

$$DF = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(B^H(\varphi_1), \dots, B^H(\varphi_n))\varphi_i.$$

The following integration-by-parts formula holds.

**Lemma 3.2.** *Let $F$ be an elementary random variable of the form* (3.8). *Then, for any $\varphi \in \mathcal{H}$ we have*

$$\mathbb{E}(\langle DF, \varphi\rangle_{\mathcal{H}}) = \mathbb{E}(FB^H(\varphi)). \tag{3.9}$$

*Proof.* First notice that we can normalize Eq. (3.9) and assume that the norm of $\varphi$ is one. There exist orthonormal elements of $\mathcal{H}$, $e_1, \dots, e_n$, such that $\varphi = e_1$ and $F$ is an elementary random variable of the form

$$F = f(B^H(e_1), \dots, B^H(e_n)),$$

where $f$ is in $C_p^\infty(\mathbb{R}^n)$. Let $\phi(x)$ denote the density of the standard normal distribution on $\mathbb{R}^n$, that is,

$$\phi(x) = (2\pi)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\Big).$$

Then we have

$$\begin{aligned}\mathbb{E}(\langle DF, \varphi\rangle_{\mathcal{H}}) &= \int_{\mathbb{R}^n} \frac{\partial f}{\partial x_1}(x)\phi(x)\,dx \\ &= \int_{\mathbb{R}^n} f(x)\phi(x)x_1\,dx \\ &= \mathbb{E}(FB^H(e_1)) = \mathbb{E}(FB^H(\varphi)),\end{aligned}$$

which completes the proof of the lemma. □

As a consequence, if $F$ and $G$ are elementary random variables and $h \in \mathcal{H}$, then we have

$$\mathbb{E}(G\langle DF, h\rangle_{\mathcal{H}}) = \mathbb{E}(-F\langle DG, h\rangle_{\mathcal{H}} + FGB^H(h)). \tag{3.10}$$

Formula (3.10) implies that the derivative operator $D$ is a closable operator from $L^p(\Omega)$ into $L^p(\Omega; \mathcal{H})$, for any $p \geq 1$. We denote by he Sobolev space $\mathbb{D}^{1,p}$ is the closure of $\mathcal{S}$ with respect to the norm

$$\|F\|_{1,p} = [\mathbb{E}(|F|^p) + \mathbb{E}(\|DF\|^p_{\mathcal{H}})]^{1/p}.$$

One can interpret $\mathbb{D}^{1,p}$ as an infinite-dimensional weighted Sobolev space.

The *divergence operator* $\delta$ is the adjoint of the derivative operator. That is, we say that a random variable $u$ in $L^2(\Omega; \mathcal{H})$ belongs to the domain of the divergence operator, denoted by Dom $\delta$, if

$$|\mathbb{E}(\langle DF, u\rangle_{\mathcal{H}})| \leq c_u \|F\|_{L^2(\Omega)}$$

for any $F \in \mathcal{S}$. In this case $\delta(u)$ is defined by the duality relationship

$$\mathbb{E}(F\delta(u)) = \mathbb{E}(\langle DF, u\rangle_{\mathcal{H}}), \tag{3.11}$$

for any $F \in \mathbb{D}^{1,2}$.

For example, consider an elementary $\mathcal{H}$-valued random variable of the form $u = \sum_{k=1}^m F_k \varphi_k$, where $F_k \in \mathbb{D}^{1,2}$ and $\varphi_k \in \mathcal{H}$. Then, $u$ belongs to the domain of the divergence and from (3.10) we deduce

$$\delta(u) = \sum_{k=1}^m [F_k B^H(\varphi_k) - \langle DF_k, \varphi_k\rangle_{\mathcal{H}}]. \tag{3.12}$$

The expression $F_k B^H(\varphi_k) - \langle DF_k, \varphi_k\rangle_{\mathcal{H}}$ is called the *Wick product* of the random variables $F_k$ and $B^H(\varphi_k)$ and it is denoted by

$$F_k \diamond B^H(\varphi_k) = F_k B^H(\varphi_k) - \langle DF_k, \varphi_k\rangle_{\mathcal{H}}. \tag{3.13}$$

With this notation (3.12) can be written as

$$\delta(u) = \sum_{k=1}^m F_k \diamond B^H(\varphi_k).$$

We will make use of the notation

$$\delta(u) = \int_0^T u_t \diamond dB^H_t,$$

when $u$ is a stochastic process in the domain of the divergence operator.

Here are some basic formulas of the Malliavin calculus which hold for any elementary random variables $F$ and $u$.

$$\mathbb{E}(\delta(u)^2) = \mathbb{E}(\|u\|^2_{\mathcal{H}}) + \mathbb{E}(\langle Du, (Du)^*\rangle_{\mathcal{H}\otimes\mathcal{H}}), \tag{3.14}$$

$$\delta(Fu) = F\delta(u) - \langle DF, u\rangle_{\mathcal{H}}, \tag{3.15}$$

$$\langle D(\delta(u)), h\rangle_{\mathcal{H}} = \langle u, h\rangle_{\mathcal{H}} + \delta(\langle Du, h\rangle_{\mathcal{H}}), \tag{3.16}$$

where $(Du)^*$ is the adjoint of $Du$ in the Hilbert space $\mathcal{H} \otimes \mathcal{H}$. Equation (3.14) holds for any $u$ in the Sobolev space $\mathbb{D}^{1,2}(\mathcal{H})$ of $\mathcal{H}$-valued random variables and it implies that $\mathbb{D}^{1,2}(\mathcal{H}) \subset \operatorname{Dom} \delta$ . Equation (3.15) holds if $F \in \mathbb{D}^{1,2}$, $u$ belongs to the domain of $\delta$ and $Fu$ and $F\delta(u) + \langle DF, u\rangle_{\mathcal{H}}$ are square integrable. Finally, the commutation relation (3.16) holds for any $h \in \mathcal{H}$ and $u \in \mathbb{D}^{1,2}(\mathcal{H})$ such that $\delta(u) \in \mathbb{D}^{1,2}$:

In case of an ordinary Brownian motion, the adapted processes in $L^2([0, T] \times \Omega)$ belong to the domain of the divergence operator, and on this class of processes the divergence operator coincides with the Itô stochastic integral (see Nualart and Pardoux [34]). Actually, the divergence operator coincides with an extension of Itô's stochastic integral introduced by Skorohod in [42]. This is a consequence of formula (3.13), because if $\varphi_k = \mathbf{1}_{[a_k,b_k]}$ and $F_k$ is a random variable measurable with respect to the $\sigma$-field generated by $\{B_t^{1/2}, t \le a_k\}$, then $\langle DF_k, \mathbf{1}_{[a_k,b_k]}\rangle_{L^2([0,T])} = 0$, and the Wick product of $F_k$ and $B_{b_k}^{1/2} - B_{a_k}^{1/2}$ is equal to the ordinary product. Notice here that the random variables $F_k$ and $B_{b_k}^{1/2} - B_{a_k}^{1/2}$ are independent.

**3.4. Wick integrals with respect to the fBm.** A natural question in this framework is to ask in which sense the divergence operator with respect to a fractional Brownian motion $B$ can be interpreted as a stochastic integral. The following proposition provides an answer to this question.

**Proposition 3.3.** *Fix a time interval* $[0, T]$. *Let* $F$ *be a function of class* $C^1$ *such which satisfies, together with* $F'$, *the growth condition*

$$|F(x)| \le ce^{\lambda x^2}, \tag{3.17}$$

*where* $c$ *and* $\lambda$ *are positive constants such that* $\lambda < \frac{1}{4T^{2H}}$. *Suppose that* $H > \frac{1}{2}$. *Then,* $F(B_t^H)$ *belongs to the domain of the divergence operator and*

$$\int_0^T F(B_t^H) \diamond dB_t^H = \int_0^T F(B_t^H)\, dB_t^H - H \int_0^T F'(B_t^H) t^{2H-1}\, dt, \tag{3.18}$$

*where* $\int_0^T F(B_t^H)\, dB_t^H$ *is the path-wise Riemann–Stieltjes integral.*

**Remarks.** 1. Formula (3.18) leads to the following equation for the expectation of a path-wise integral:

$$\mathbb{E}\left( \int_0^T F(B_t^H)\, dB_t^H \right) = H \int_0^T \mathbb{E}(F'(B_t^H)) t^{2H-1}\, dt.$$

2. Suppose that $F$ is a function of class $C^2$ such that $F$, $F'$ and $F''$ satisfy the growth condition (3.17). Then, (3.18) and (3.7) yield

$$F(B_T^H) = F(0) + \int_0^T F'(B_t^H) \diamond dB_t^H + H \int_0^T F''(B_t^H) t^{2H-1}\, dt, \tag{3.19}$$

which can be considered as an Itô formula for the Wick integral.

*Proof of Proposition* 3.3. Set $t_i = \frac{iT}{n}$. Then formula (3.13) yields

$$\sum_{i=1}^{n} F(B^H_{t_{i-1}}) \diamond (B^H_{t_i} - B^H_{t_{i-1}}) = \sum_{i=1}^{n} F(B^H_{t_{i-1}})(B^H_{t_i} - B^H_{t_{i-1}}) - \sum_{i=1}^{n} \langle D(F(B^H_{t_{i-1}})), \mathbf{1}_{[t_{i-1},t_i]} \rangle_{\mathcal{H}}.$$

We have, using the chain rule and $DB^H_{t_{i-1}} = \mathbf{1}_{[0,t_{i-1}]}$,

$$\begin{aligned} \langle D(F(B^H_{t_{i-1}})), \mathbf{1}_{[t_{i-1},t_i]} \rangle_{\mathcal{H}} &= F'(B^H_{t_{i-1}}) \langle \mathbf{1}_{[0,t_{i-1}]}, \mathbf{1}_{[t_{i-1},t_i]} \rangle_{\mathcal{H}} \\ &= F'(B^H_{t_{i-1}})(R_H(t_{i-1}, t_i) - R_H(t_{i-1}, t_{i-1})) \\ &= \frac{1}{2} F'(B^H_{t_{i-1}})((t_i)^{2H} - (t_{i-1})^{2H} - (t_i - t_{i-1})^{2H}). \end{aligned}$$

Then it suffices to take the limit as $n$ tends to infinity. The convergences are almost surely and in $L^2(\Omega)$. □

As an application of Proposition 3.3 we will derive the following estimate for the variance of the path-wise stochastic integral of a trigonometric function.

**Proposition 3.4.** *Let $B^H$ be a $d$-dimensional fractional Brownian motion with Hurst parameter $H > 1/2$. Then for any $\xi \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left( \left\| \int_0^T e^{i\langle \xi, B^H_t \rangle} \, dB^H_t \right\|^2_{\mathbb{C}} \right) \leq C(1 \wedge |\xi|^{\frac{1}{H}-2}), \tag{3.20}$$

*where $\|z\|_{\mathbb{C}} = \sum_{i=1}^d z^i \bar{z}^i$ and $C$ is a constant depending on $T$, $d$ and $H$.*

*Proof.* From (3.18) we get

$$\int_0^T e^{i\langle \xi, B^H_t \rangle} \, dB^H_t = \int_0^T e^{i\langle \xi, B^H_t \rangle} \diamond dB^H_t + H \int_0^T i\xi e^{i\langle \xi, B^H_t \rangle} t^{2H-1} \, dt. \tag{3.21}$$

We denote by $\pi_\xi(x) = x - \frac{\xi}{|\xi|^2}\langle \xi, x \rangle$ the projection operator on the orthogonal subspace of $\xi$. Clearly

$$\int_0^T e^{i\langle \xi, B^H_t \rangle} \, dB^H_t = \pi_\xi\left( \int_0^T e^{i\langle \xi, B^H_t \rangle} \, dB^H_t \right) + \frac{i\xi}{|\xi|^2}(e^{i\langle \xi, B^H_T \rangle} - 1), \tag{3.22}$$

and, as a consequence, it suffices to show the estimate (3.20) for the first summand in the right-hand side of (3.22). From (3.21) it follows that

$$Z := \pi_\xi\left( \int_0^T e^{i\langle \xi, B^H_t \rangle} \, dB^H_t \right) = \pi_\xi\left( \int_0^T e^{i\langle \xi, B^H_t \rangle} \diamond dB^H_t \right).$$

Then we need to compute the expectation of the square norm of the $\mathbb{C}^3$-valued random variable $Z$. This is done using the duality relationship (3.11) and the commutation formula (3.16). The composition of the projection operator $\pi_\xi$ and the derivative operator $D$ vanishes on a random variable of the form $e^{i\langle \xi, B_t^H\rangle}$. Hence, only the first term in the commutation formula (3.16) applied to $u_t = e^{i\langle \xi, B_t^H\rangle}$ will contribute to $\mathbb{E}(\|Z\|_{\mathbb{C}}^2)$ and we obtain

$$\begin{aligned}
\mathbb{E}(\|Z\|_{\mathbb{C}}^2) &= \sum_{j=1}^{d} \mathbb{E}(Z^j \overline{Z^j}) \\
&= \sum_{j=1}^{d} \left(1 - \frac{(\xi^j)^2}{|\xi|^2}\right) \mathbb{E}\big(\big\langle e^{-i\langle \xi, B_\cdot^H\rangle}, e^{-i\langle \xi, B_\cdot^H\rangle}\big\rangle_{\mathcal{H}}\big) \\
&= (d-1)\alpha_H \int_0^T \int_0^T \mathbb{E}\big(e^{i\langle \xi, B_s^H - B_r^H\rangle}\big)|s-r|^{2H-2}\, ds dr \\
&= (d-1)\alpha_H \int_0^T \int_0^T e^{-\frac{|s-r|^{2H}}{2}|\xi|^2}|s-r|^{2H-2}\, ds dr,
\end{aligned}$$

which leads to the desired estimate. □

Proposition 3.3 also holds for $H \in \left(\frac{1}{4}, \frac{1}{2}\right]$ if we replace the path-wise integral in the right-hand side of (3.18) by the *Stratonovich integral* defined as the limit in probability of symmetric sums

$$\int_0^T F(B_t^H)\, dB_t^H = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{2}\big[F\big(B^H_{\frac{(i-1)T}{n}}\big) + F\big(B^H_{\frac{iT}{n}}\big)\big]\big(B^H_{\frac{iT}{n}} - B^H_{\frac{(i-1)T}{n}}\big).$$

For $H = 1/2$ the Wick integral appearing in Equation (3.18) is the classical Itô integral and it is the limit of forward Wick or ordinary Riemann sums:

$$\begin{aligned}
\int_0^T F\big(B_t^{1/2}\big) \diamond dB_t^{1/2} &= \lim_{n\to\infty} \sum_{i=1}^{n} F\big(B^{1/2}_{\frac{(i-1)T}{n}}\big) \diamond \big(B^{1/2}_{\frac{iT}{n}} - B^{1/2}_{\frac{(i-1)T}{n}}\big) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} F\big(B^{1/2}_{\frac{(i-1)T}{n}}\big)\big(B^{1/2}_{\frac{iT}{n}} - B^{1/2}_{\frac{(i-1)T}{n}}\big).
\end{aligned}$$

Nevertheless, for $H < 1/2$ the forward Riemann sums do not converge in general. For example, in the simplest case $F(x) = x$, we have, with the notation $t_i = \frac{iT}{n}$

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^{n} (B^H_{t_{i-1}}(B^H_{t_i} - B^H_{t_{i-1}}))\Big) &= \frac{1}{2}\sum_{i=1}^{n}\big[t_i^{2H} - t_{i-1}^{2H} - (t_i - t_{i-1})^{2H}\big] \\
&= \frac{1}{2}T^{2H}(1 - n^{1-2H}) \to -\infty,
\end{aligned}$$

as $n$ tends to infinity.

The convergence of the forward Wick Riemann sums to the forward Wick integral in the case $H \in \left(\frac{1}{4}, \frac{1}{2}\right)$ has been recently established in [36] and [5]. More precisely, the following theorem has been proved in [36].

**Theorem 3.5.** *Suppose $H \in \left(\frac{1}{4}, \frac{1}{2}\right)$ and let $F$ be a function of class $C^7$ such that $F$ together with its derivatives satisfy the growth condition* (3.17). *Then, the forward Wick integral*

$$\int_0^T F(B_t^H) \diamond dB_t^H = \lim_{n\to\infty} \sum_{i=1}^n F\left(B^H_{\frac{(i-1)T}{n}}\right) \diamond \left(B^H_{\frac{iT}{n}} - B^H_{\frac{(i-1)T}{n}}\right)$$

*exists and the Wick–Itô formula* (3.19) *holds.*

More generally, we can replace the fractional Brownian motion $B^H$ by an arbitrary Gaussian process $\{X_t, t \geq 0\}$ with zero mean and continuous covariance function $R(s,t) = \mathbb{E}(X_s X_t)$. Suppose that the variance function $V_t = \mathbb{E}(X_t^2)$ has bounded variation on any finite interval and the following conditions hold for any $T > 0$:

$$\lim_{n\to\infty} \sum_{i,j=1}^n \left(\mathbb{E}\left(\left(X_{\frac{iT}{n}} - X_{\frac{(i-1)T}{n}}\right)\left(X_{\frac{jT}{n}} - X_{\frac{(j-1)T}{n}}\right)\right)\right)^2 \to 0, \tag{3.23}$$

$$\lim_{n\to\infty} \sum_{i=1}^n \sup_{0\leq t\leq T} \left(\mathbb{E}\left(\left(X_{\frac{iT}{n}} - X_{\frac{(i-1)T}{n}}\right)X_t\right)\right)^2 \to 0. \tag{3.24}$$

Then it is proved in [36] that the forward Wick integral $\int_0^T F(X_t) \diamond dX_t$ exists and the following version of the Wick–Itô formula holds:

$$F(X_T) = F(X_0) + \int_0^T F'(X_t) \diamond dX_t + \frac{1}{2}\int_0^T F''(X_t)\, dV_t.$$

## 4. Application of fBm in turbulence

The observations of three-dimensional turbulent fluids indicate that the vorticity field of the fluid is concentrated along thin structures called vortex filaments. In his book Chorin [10] suggests probabilistic descriptions of vortex filaments by trajectories of self-avoiding walks on a lattice. Flandoli [17] introduced a model of vortex filaments based on a three-dimensional Brownian motion. A basic problem in these models is the computation of the kynetic energy of a given configuration.

Denote by $u(x)$ the velocity field of the fluid at point $x \in \mathbb{R}^3$, and let $\xi = \mathrm{curl} u$ be the associated vorticity field. The kynetic energy of the field will be

$$\mathbb{H} = \frac{1}{2}\int_{\mathbb{R}^3} |u(x)|^2 dx = \frac{1}{8\pi}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\xi(x)\cdot\xi(y)}{|x-y|}\, dxdy. \tag{4.1}$$

We will assume that the vorticity field is concentrated along a thin tube centered in a curve $\gamma = \{\gamma_t, 0 \le t \le T\}$. Moreover, we will choose a random model and consider this curve as the trajectory of a three-dimensional fractional Brownian motion $B^H = \{B_t^H, 0 \le t \le T\}$ with Hurst parameter $H$. That is, the components of the process $B^H$ are independent fractional Brownian motions. This modelization is justified by the fact that the trajectories of the fractional Brownian motion are Hölder continuous of any order $H \in (0, 1)$. For technical reasons we are going to consider only the case $H > \frac{1}{2}$.

Then the vorticity field can be formally expressed as

$$\xi(x) = \Gamma \int_{\mathbb{R}^3} \left( \int_0^T \delta(x - y - B_s^H) \dot{B}_s^H \, ds \right) \rho\,(dy), \tag{4.2}$$

where $\Gamma$ is a parameter called the circuitation, and $\rho$ is a probability measure on $\mathbb{R}^3$ with compact support.

Substituting (4.2) into (4.1) we derive the following formal expression for the kinetic energy:

$$\mathbb{H} = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \mathbb{H}_{xy} \rho(dx) \rho(dy), \tag{4.3}$$

where the so-called interaction energy $\mathbb{H}_{xy}$ is given by the double integral

$$\mathbb{H}_{xy} = \frac{\Gamma^2}{8\pi} \sum_{i=1}^3 \int_0^T \int_0^T \frac{1}{|x + B_t^H - y - B_s^H|} \, dB_s^{H,i} dB_t^{H,i}. \tag{4.4}$$

We are interested in the following problems: Is $\mathbb{H}$ a well defined random variable? Does it have moments of all orders and even exponential moments?

In order to give a rigorous meaning to the double integral (4.4) we introduce the regularization of the function $|\cdot|^{-1}$:

$$\sigma_n = |\cdot|^{-1} * p_{1/n}, \tag{4.5}$$

where $p_{1/n}$ is the Gaussian kernel with variance $\frac{1}{n}$. Then the smoothed interaction energy

$$\mathbb{H}_{xy}^n = \frac{\Gamma^2}{8\pi} \sum_{i=1}^3 \int_0^T \left( \int_0^T \sigma_n(x + B_t^H - y - B_s^H) \, dB_s^{H,i} \right) dB_t^{H,i} \tag{4.6}$$

is well defined, where the integrals are path-wise Riemann–Stieltjes integrals. Set

$$\mathbb{H}^n = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \mathbb{H}_{xy}^n \rho(dx) \rho(dy). \tag{4.7}$$

The following result has been proved in [35].

**Theorem 4.1.** *Suppose that the measure $\rho$ satisfies*

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} |x-y|^{1-\frac{1}{H}}\rho(dx)\rho(dy) < \infty. \tag{4.8}$$

*Let $\mathbb{H}^n_{xy}$ be the smoothed interaction energy defined by* (4.5). *Then $\mathbb{H}^n$ defined in* (4.7) *converges, for all $k \geq 1$, in $L^k(\Omega)$ to a random variable $\mathbb{H} \geq 0$ that we call the energy associated with the vorticity field* (4.2).

If $H = \frac{1}{2}$, the fBm $B^H$ is a classical three-dimensional Brownian motion. In this case condition (4.8) would be $\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} |x-y|^{-1}\rho(dx)\rho(dy) < \infty$, which is the assumption made by Flandoli [17] and Flandoli and Gubinelli [18]. In this last paper, using Fourier approach and Itô's stochastic calculus, the authors show that $\mathbb{E}(e^{-\beta\mathbb{H}}) < \infty$ for sufficiently small negative $\beta$.

The proof of Theorem 4.1 is based on the stochastic calculus with respect to fBm and the application of Fourier transform. Using Fourier transform we can write

$$\frac{1}{|z|} = \int_{\mathbb{R}^3} (2\pi)^3 \frac{e^{-i\langle\xi,z\rangle}}{|\xi|^2}\, d\xi$$

and

$$\sigma_n(x) = \int_{\mathbb{R}^3} |\xi|^{-2} e^{i\langle\xi,x\rangle - |\xi|^2/2n}\, d\xi. \tag{4.9}$$

Substituting (4.9) into in (4.6), we obtain the following formula for the smoothed interaction energy:

$$\begin{aligned}\mathbb{H}^n_{xy} &= \frac{\Gamma^2}{8\pi}\sum_{j=1}^3 \int_0^T\int_0^T \left(\int_{\mathbb{R}^3} e^{i\langle\xi,x+B_t-y-B_s\rangle}\frac{e^{-|\xi|^2/2n}}{|\xi|^2}\right) dB_s^{H,j}\, dB_t^{H,j} \\ &= \frac{\Gamma^2}{8\pi}\int_{\mathbb{R}^3} |\xi|^{-2} e^{i\langle\xi,x-y\rangle - |\xi|^2/2n}\|Y_\xi\|^2_{\mathbb{C}}\, d\xi, \end{aligned} \tag{4.10}$$

where

$$Y_\xi = \int_0^T e^{i\langle\xi, B_t^H\rangle}\, dB_t^H.$$

Integrating with respect to $\rho$ yields

$$\mathbb{H}^n = \frac{\Gamma^2}{8\pi}\int_{\mathbb{R}^3} \|Y_\xi\|^2_{\mathbb{C}} |\xi|^{-2} |\widehat{\rho}(\xi)|^2 e^{-|\xi|^2/2n} d\xi \geq 0. \tag{4.11}$$

From Fourier analysis and condition (4.8) we know that

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} |x-y|^{1-\frac{1}{H}}\rho(dx)\rho(dy) = C_H \int_{\mathbb{R}^3} |\widehat{\rho}(\xi)|^2 |\xi|^{\frac{1}{H}-4} d\xi < \infty. \tag{4.12}$$

Then, taking into account (4.12) and (4.11), in order to show the convergence in $L^k(\Omega)$ of $\mathbb{H}^n$ to a random variable $\mathbb{H} \geq 0$ it suffices to check that

$$\mathbb{E}(\|Y_\xi\|_{\mathbb{C}}^{2k}) \leq C_k\big(1 \wedge |\xi|^{k(\frac{1}{H}-2)}\big). \tag{4.13}$$

For $k = 2$ this has been proved in Proposition 3.4. The general case $k \geq 2$ follows by similar arguments making use of the *local nondeterminism property* of fBm (see Berman [4]):

$$\mathrm{Var}\Big(\sum_i (B_{t_i}^H - B_{s_i}^H)\Big) \geq k_H \sum_i (t_i - s_i)^{2H}.$$

## 5. Application to financial mathematics

Fractional Brownian motion has been used to describe the behavior to prices of assets and volatilities in stock markets. The long-range dependence self-similarity properties make this process a suitable model to describe these quantities. We refer to Shiryaev [41] for a general description of the applications of fractional Brownian motion to model financial quantities. We will briefly present in this section two different uses of fBm in mathematical finance.

**5.1. Fractional Black and Scholes model.** It has been proposed by several authors to replace the classical Black and Scholes model which has no memory and is based on the geometric Brownian motion by the so-called *fractional Black and Scholes model*. In this model the market stock price of the risky asset is given by

$$S_t = S_0 \exp\left(\mu t + \sigma B_t^H - \frac{\sigma^2}{2} t^{2H}\right), \tag{5.1}$$

where $B^H$ is an fBm with Hurst parameter $H$, $\mu$ is the mean rate of return and $\sigma > 0$ is the volatility. The price of the non-risky assets at time $t$ is $e^{rt}$, where $r$ is the interest rate.

Consider an investor who starts with some initial endowment $V_0 \geq 0$ and invests in the assets described above. Let $\alpha_t$ be the number of non-risky assets and let $\beta_t$ the number of stocks owned by the investor at time $t$. The couple $(\alpha_t, \beta_t)$, $t \in [0, T]$ is called a *portfolio* and we assume that $\alpha_t$ and $\beta_t$ are stochastic processes. Then the investor's wealth or value of the portfolio at time $t$ is

$$V_t = \alpha_t e^{rt} + \beta_t S_t.$$

We say that the portfolio is *self-financing* if

$$V_t = V_0 + r \int_0^t \alpha_s e^{rs} ds + \int_0^t \beta_s dS_s. \tag{5.2}$$

This means that there is no fresh investment and there is no consumption. We see here that the self-financing condition requires the definition of a stochastic integral with respect to the fBm, and there are two possibilities: path-wise integrals and Wick-type integrals.

The use of path-wise integrals leads to the existence of arbitrage opportunities, which is one of the main drawbacks of the model (5.1). Different authors have proved the existence of arbitrages for the fractional Black and Scholes model (see Rogers [40], Shiryaev [41], and Cheridito [9]). By definition, an arbitrage is a self-financing portfolio which satisfies $V_0 = 0$, $V_T \geq 0$ and $P(V_T > 0) > 0$.

In the case $H > \frac{1}{2}$, one can construct an arbitrage in the following simple way. Suppose, to simplify, that $\mu = r = 0$. Consider the self-financing portfolio defined by

$$\begin{aligned} \beta_t &= S_t - S_0, \\ \alpha_t &= \int_0^t \beta_s dS_s - \beta_t S_t. \end{aligned}$$

This portfolio satisfies $V_0 = 0$ and $V_t = (S_t - S_0)^2 > 0$ for all $t > 0$, and hence it is an arbitrage.

In the classical Black and Scholes model (case $H = \frac{1}{2}$), there exists an equivalent probability measure $Q$ under which $\mu = r$ and the discounted price process $\widetilde{S}_t = e^{-rt} S_t$ is a martingale. Then, the discounted value of a self-financing adapted portfolio satisfying $\mathbb{E}_Q\big(\int_0^T \beta_s^2 \widetilde{S}_s^2 ds\big) < \infty$ is a martingale on the time interval $[0, T]$ given by the Itô stochastic integral

$$\widetilde{V}_t = V_0 + \int_0^t \beta_s d\widetilde{S}_s.$$

As a consequence, $V_t = e^{-r(T-t)} \mathbb{E}_Q(V_T | \mathcal{F}_t)$, and the price of an European option with payoff $G$ at the maturity time $T$ is given by $e^{-r(T-t)} \mathbb{E}_Q(G|\mathcal{F}_t)$. The probability $Q$ is called the martingale measure. In the case $H \neq \frac{1}{2}$, there exist an equivalent probability $Q$ under which $\mu = r$ and $S_t = S_0 \exp(\sigma B_t^H - \frac{\sigma^2}{2} t^{2H})$ has constant expectation. However, $e^{-rt} S_t$ is not a martingale under $Q$.

The existence of arbitrages can be avoided using forward Wick integrals to define the self-financing property (5.2). In fact, using the Wick–Itô formula in (5.1) yields

$$dS_t = \mu S_t dt + \sigma S_t \diamond dB_t^H,$$

and then the self-financing condition (5.2) could be written as

$$V_t = V_0 + \int_0^t (r\alpha_s e^{rs} + \mu \beta_s S_s) ds + \sigma \int_0^t \beta_s S_s \diamond dB_s^H.$$

Applying the stochastic calculus with respect to the Wick integral, Hu and Øksendal in [22], and Elliott and Hoek in [16] have derived the following formula for the value

of the call option with payoff $(S_T - K)^+$ at time $t \in [0, T]$:

$$C(t, S_t) = S_t \Phi(y_+) - K e^{-r(T-t)} \Phi(y_-), \tag{5.3}$$

where

$$y_\pm = \left( \ln \frac{S_t}{K} + r(T-t) \pm \frac{\sigma^2 (T^{2H} - t^{2H})}{2} \right) \Big/ \sigma \sqrt{T^{2H} - t^{2H}}. \tag{5.4}$$

In [6] Björk and Hult argue that the definition of a self-financing portfolio using the Wick product is quite restrictive and in [37] Nualart and Taqqu explain the fact that in formula (5.4) only the increment of the variance of the process in the interval $[t, T]$ appears, and extend this formula to price models driven by a general Gaussian process.

**5.2. Stochastic volatility models.** It has been observed that in the classical Black and Scholes model the implied volatility $\sigma^{\text{imp}}_{t,T}$ obtained from formula (5.3) for different options written on the same asset is not constant and heavily depends on the time $t$, the time to maturity $T - t$ and the strike price $S_t$. The $U$-shaped pattern of implied volatilites across different strike prices is called "smile", and it is believed that this and other features as the volatility clustering can be explained by stochastic volatility models. Hull and White have proposed in [23] an option pricing model in which the volatility of the asset price is of the form $\exp(Y_t)$, where $Y_t$ is an Ornstein–Uhlenbeck process.

Consider the following stochastic volatility model based on the fractional Ornstein–Uhlenbeck process. The price of the asset $S_t$ is given by

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t,$$

where $\sigma_t = f(Y_t)$ and $Y_t$ is a fractional Ornstein–Uhlenbeck process:

$$dY_t = \alpha(m - Y_t)dt + \beta_t dB_t^H.$$

The process $W_t$ is an ordinary Brownian motion and $B_t^H$ is a fractional Brownian motion with Hurst parameter $H > \frac{1}{2}$, independent of $W$. Examples of functions $f$ are $f(x) = e^x$ and $f(x) = |x|$.

Comte and Renault studied in [11] this type of stochastic volatility model which introduces long memory and mean reverting in the Hull and White setting. The long-memory property allows this model to capture the well-documented evidence of persistence of the stochastic feature of Black and Scholes implied volatilities, when time to maturity increases.

Hu has proved in [21] the following properties of this model.

1) The market is incomplete and martingale measures are not unique.

2) Set $\gamma_t = (r - \mu)/\sigma_t$ and

$$\frac{dQ}{dP} = \exp\left( \int_0^T \gamma_t \, dW_t - \frac{1}{2} \int_0^T |\gamma_t|^2 \, dt \right).$$

Then $Q$ is the *minimal martingale measure* associated with $P$.

3) The risk minimizing-hedging price at time $t = 0$ of an European call option with payoff $(S_T - K)^+$ is given by

$$C_0 = e^{-rT}\mathbb{E}_Q[(S_T - K)^+]. \tag{5.5}$$

As a consequence of (5.5), if $\mathcal{G}_t$ denotes the filtration generated by fBm, we obtain

$$\begin{aligned} C_0 &= e^{-rT}\mathbb{E}_Q[\mathbb{E}_Q((S_T - K)^+|\mathcal{G}_T)] \\ &= e^{-rT}\mathbb{E}_Q[C_{BS}(\sigma)]. \end{aligned}$$

Here $\sigma = \sqrt{\int_0^T \sigma_s^2 ds}$ and $C_{BS}(\sigma)$ is the Black and Scholes price function given by

$$C_{BS} = S_0\Phi(y_+) - Ke^{-rT}\Phi(y_-),$$

where

$$y_\pm = \frac{\ln\frac{S_0}{K} + (r \pm \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}.$$

## References


[1] Alòs, E., Mazet, O., Nualart, D., Stochastic calculus with respect to fractional Brownian motion with Hurst parameter lesser than $\frac{1}{2}$. *Stoch. Proc. Appl.* **86** (1999), 121–139.

[2] Alòs, E., Mazet, O., Nualart, D., Stochastic calculus with respect to Gaussian processes. *Ann. Probab.* **29** (2001), 766–801.

[3] Alòs, E., Nualart, D., Stochastic integration with respect to the fractional Brownian motion. *Stoch. Stoch. Rep.* **75** (2003), 129–152.

[4] Berman, S., Local nondeterminism and local times of Gaussian processes. *Indiana Univ. Math. J.* **23** (1973), 69–94.

[5] Biagini, F., Øksendal, B., Forward integrals and an Itô formula for fractional Brownian motion. Preprint, 2005.

[6] Björk, R., Hult, H., A note on Wick products and the fractional Black-Scholes model. Preprint, 2005.

[7] Carmona, P. , Coutin, L., Stochastic integration with respect to fractional Brownian motion. *Ann. Inst. Henri Poincaré* **39** (2003), 27–68.

[8] Cheridito, P., Mixed fractional Brownian motion. *Bernoulli* **7** (2001), 913–934.

[9] Cheridito, P., Regularizing Fractional Brownian Motion with a View towards Stock Price Modelling. PhD Dissertation, ETH, Zürich, 2001.

[10] Chorin, A., *Vorticity and Turbulence*. Appl. Math. Sci. 103, Springer-Verlag, New York 1994.

[11] Comte, F., Renault, E., Long memory in continuous-time stochastic volatility models. *Math. Finance* **8** (1998), 291–323

[12] Coutin, L., Qian, Z., Stochastic analysis, rough paths analysis and fractional Brownian motions. *Probab. Theory Related Fields* **122** (2002), 108–140.

[13] Dai, W., Heyde, C. C., Itô's formula with respect to fractional Brownian motion and its application. *J. Appl. Math. Stochastic Anal.* **9** (1996), 439–448.

[14] Decreusefond, L., Üstünel, A. S., Stochastic analysis of the fractional Brownian motion. *Potential Anal.* **10** (1998), 177–214.

[15] Duncan, T. E., Hu, Y., Pasik-Duncan, B., Stochastic calculus for fractional Brownian motion I. Theory. *SIAM J. Control Optim.* **38** (2000), 582–612.

[16] Elliott, R. J., van der Hoek, J., A general fractional white noise theory and applications to finance. *Math. Finance* **13** (2003), 301–330.

[17] Flandoli, F., On a probabilistic description of small scale structures in 3D fluids. *Ann. Inst. Henri Poincaré* **38** (2002), 207–228.

[18] Flandoli, F., Gubinelli, M., The Gibbs ensemble of a vortex filament. *Probab. Theory Related Fields* **122** (2001), 317–340.

[19] Garsia, A. M., Rodemich, E., Rumsey, H., Jr., A real variable lemma and the continuity of paths of some Gaussian processes. *Indiana Univ. Math. J.* **20** (1970/1971), 565–578.

[20] Guerra, J., Nualart, D., The $1/H$-variation of the divergence integral with respect to the fractional Brownian motion for $H > 1/2$ and fractional Bessel processes. *Stoch. Process. Appl.* **115** (2005), 91–115.

[21] Hu, Y., Integral transformations and anticipative calculus for fractional Brownian motions. *Mem. Amer. Math. Soc.* **175** (2005).

[22] Hu, Y. , Øksendal, B., Fractional white noise calculus and applications to finance. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* **6** (2003), 1–32.

[23] Hull, J., White, A., The pricing of options on assets with stochastic volatilities. *J. Finance* **3** (1987), 281–300.

[24] Hurst, H., E. Long-term storage capacity in reservoirs. *Trans. Amer. Soc. Civil Eng.* **116** (1951), 400–410.

[25] Kolmogorov, A. N., Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum. *C. R.* (*Doklady*) *Acad. URSS* (*N.S.*) **26** (1940), 115–118.

[26] Lin, S. J., Stochastic analysis of fractional Brownian motions. *Stoch. Stoch. Rep.* **55** (1995), 121–140.

[27] Lyons, T., Differential equations driven by rough signals. *Rev. Mat. Iberoamericana* **14** (1998), 215–310.

[28] Lyons, T., Qian, Z., *System control and rough paths*. Oxford Math. Monogr., Oxford University Press, Oxford 2002.

[29] Malliavin, P. , Stochastic calculus of variations and hypoelliptic operators. In *Proceedings of the International Symposium on Stochastic Differential Equations* (Kyoto, 1976), Wiley, New York, Chichester, Brisbane 1978, 195–263.

[30] Mandelbrot, B. B., Van Ness, J. W., Fractional Brownian motions, fractional noises and applications. *SIAM Review* **10** (1968), 422–437.

[31] Memin, J., Mishura, Y., Valkeila, E., Inequalities for the moments of Wiener integrals with respecto to fractional Brownian motions. *Statist. Prob. Letters* **55** (2001), 421–430.

[32] Nualart, D., *The Malliavin calculus and related topics.* 2nd edition, Probab. Appl., Springer Verlag, New York 2005.

[33] Nualart, D., Stochastic integration with respect to fractional Brownian motion and applications. *Contemp. Math.* **336** (2003), 3–39.

[34] Nualart, D., Pardoux, E., Stochastic calculus with anticipating integrands. *Probab. Theory Related Fields* **78** (1988), 535–581.

[35] Nualart, D., Rovira, C., Tindel, S., Probabilistic models for vortex filaments based on fractional Brownian motion. *Ann. Probab.* **31** (2003), 1862–1899.

[36] Nualart, D., Taqqu, M. S., Wick-Itô formula for Gaussian processes. *Stoch. Anal. Appl.*, to appear.

[37] Nualart, D., Taqqu, M. S., Some issues concerning Wick integrals and the Black and Scholes formula. Preprint.

[38] Pipiras, V., Taqqu, M. S., Integration questions related to fractional Brownian motion. *Probab. Theory Related Fields* **118** (2000), 121–291.

[39] Pipiras, V., Taqqu, M. S., Are classes of deterministic integrands for fractional Brownian motion on a interval complete? *Bernoulli* **7** (2001), 873–897.

[40] Rogers, L. C. G., Arbitrage with fractional Brownian motion. *Math. Finance* **7** (1997), 95–105.

[41] Shiryaev, A. N., *Essentials of Stochastic Finance: Facts, Models and Theory.* Adv. Ser. Stat. Sci. Appl. Probab. 3, World Scientific, Singapore 1999.

[42] Skorohod, A. V., On a generalization of a stochastic integral. *Theory Probab. Appl.* **20** (1975), 219–233.

[43] Young, L. C., An inequality of the Hölder type connected with Stieltjes integration. *Acta Math.* **67** (1936), 251–282.

Department of Mathematics, University of Kansas, Lawrence, Kansas 66045, U.S.A.
E-mail: nualart@math.ku.edu

# Atomistic and continuum models for phase change dynamics

Anders Szepessy

**Abstract.** The dynamics of dendritic growth of a crystal in an undercooled melt is determined by macroscopic diffusion-convection of heat and capillary forces acting on length scales compared to the nanometer width of the solid-liquid interface. Its modeling is useful for instance in processing techniques based on casting. The phase field method is widely used to study evolution of such microstructures of phase transformations on a continuum level; it couples the energy equation to a phenomenological Allen–Cahn/Ginzburg–Landau equation modeling the dynamics of an order parameter determining the solid and liquid phases, including also stochastic fluctuations to obtain the qualitative correct result of dendritic side branching. This lecture presents some ideas to derive stochastic phase field models from atomistic formulations by coarse-graining molecular dynamics and kinetic Monte Carlo methods.

**Mathematics Subject Classification (2000).** 82C31, 65C30.

**Keywords.** Phase transformation, phase-field, coarse-grained, molecular dynamics, Brownian dynamics, Langevin equation, Smoluchowski equation, kinetic Monte Carlo.

## 1. Introduction to phase-field models

The phase field model for modeling a liquid solid phase transformation is an Allen–Cahn/Ginzburg–Landau equation coupled to the energy equation

$$\begin{aligned} \partial_t \phi &= \operatorname{div}(k_1 \nabla \phi) - k_0 (f'(\phi) + g'(\phi) k_4 T) + \text{noise}, \\ \partial_t (c_v T + k_2 g(\phi)) &= \operatorname{div}(k_3 \nabla T) \end{aligned} \tag{1.1}$$

with a double-well potential $f$ having local minima at $\pm 1$, smoothed step function $g$, temperature $T$ and specific heat $c_v$, cf. [3]. The phase field variable $\phi \colon \mathbb{R}^d \times [0, \infty) \to [-1, 1]$ interprets the solid and liquid phases as the domains $\{x \in \mathbb{R}^d : \phi(x) > 0\}$ and $\{x \in \mathbb{R}^d : \phi(x) < 0\}$ respectively. To have such an implicit definition of the phases, as in the level set method, is a computational advantage compared to a sharp interface model, where the necessary direct tracking of the interface introduce computational drawbacks. This phenomenological phase-field model, with free energy potentials motived by thermodynamics, has therefore become a popular and effective computational method to solve problems with complicated microstructures of dendrite and eutectic growth, cf. [1], [3]. The phase-field model has mathematical wellposedness and convergence to sharp interface results [34].

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Assuming that the reaction term in the Allen–Cahn equation takes a given form, e.g. a standard choice is

$$f(\phi) := (1-\phi^2)^2,$$
$$g(\phi) := \tfrac{15}{16}\left(\tfrac{1}{5}\phi^5 - \tfrac{2}{3}\phi^3 + \phi\right) + \tfrac{1}{2},$$

then the parameters $k_0$, $k_1$, $k_2$, $k_3$, $k_4$ in the phase-field model can be determined from atomistic molecular simulations [19]; an alternative in [1] uses a steeper step function $g$ to easy derive consistency with sharp interface models. The evolution of the phase interface depends on the orientation of the solid crystal; this is modeled by an anisotropic matrix $k_1$. Added noise to system (1.1) is also important, e.g. to obtain sidebranching dendrites [22] explained in Section 5.4.

Phase changes can be modeled on an atomistic level by molecular dynamics or kinetic Monte Carlo methods. This lecture first presents some ideas and questions to derive a stochastic phase field model by coarse-graining molecular dynamics, to determine the reaction term (i.e. $f$ and $g$) and the noise. This is made in three steps in Sections 2 to 4: to give a precise quantitative atomistic definition of the phase-field variable, to introduce an atomistic molecular dynamics model based on Brownian dynamics, and to derive the dynamics for the coarse-grained phase-field. Section 5 derives stochastic hydrodynamical limits of solutions to an Ising model with long range interaction, i.e. coarse-graining a kinetic Monte Carlo method following [24]. Section 5.4 presents a simple kinetic Monte Carlo method for dendrite dynamics.

## 2. Quantitative atomistic definition of the phase-field variable

The aim is to give a unique definition of the phase-field variable, so that it can be determined precisely from atomistic simulations. The usual interpretation is to measure interatomic distances and use structure functions (or similar methods) to measure where the phase is solid and where it is liquid, which then implicitly defines the phase-field variable [3]. Here we instead use the energy equation for a quantitative and explicit definition of the phase-field variable. The macroscopic energy equation with a phase transformation and heat conduction is

$$\partial_t(c_v T + m) = \operatorname{div}(k\nabla T) \tag{2.1}$$

where $m$ corresponds to the latent heat release. In (1.1) the latent heat determines the parameter $k_2$, since $\phi$ is defined to jump from 1 to $-1$ in the phase transformation. We will instead use this latent heat to directly define the phase field function, and not only the parameter $k_2$. The total energy, $c_v T + m$, can be defined from molecular dynamics of $N$ particles with position $X_i$, velocity $v_i$ and mass $\mu$ in a potential $V$, see [20], [18],

$$c_v T + m = \sum_{i=1}^{N} \mu \frac{|v_i|^2}{2} + V(X_1, \ldots, X_N). \tag{2.2}$$

Assume that the potential can be defined from pair interactions

$$V(X) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}\Phi(X_i - X_j), \tag{2.3}$$

where $\Phi\colon \mathbb{R}^3 \to \mathbb{R}$ is a molecular dynamics pair potential, e.g. a Lennard–Jones potential

$$\Phi(x) = z_1\Big(\frac{\sigma}{|x|}\Big)^{12} - z_2\Big(\frac{\sigma}{|x|}\Big)^{6}.$$

In the macroscopic setting the jump of $m$ in a phase change is called the latent heat, which depends on the thermodynamic variables kept constant: with constant $N, T$ and volume it is called the internal energy and with constant pressure instead of volume it is called enthalpy. The kinetic energy $\sum_i \mu|v_i|^2/2$ is related to the temperature. It is therefore natural to let the phase field variable be determined by the potential energy $V(X)$. In a pointwise setting the potential energy can be represented by the distribution

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}\Phi(X_i - X_j)\delta(x - X_i)$$

where $\delta$ is the point mass at the origin [20]. We seek an averaged variant and we will study a microscopic phase change model where the interface is almost planar in the microscopic scale with normal in the $x_1$ direction. Therefore we take a smooth average and define the phase-field variable by

$$m(X, x) := \frac{1}{2}\sum_{i=1}^{N}\underbrace{\sum_{j\neq i}\Phi(X_i - X_j)}_{m_i(X)}\eta(x - X_i) \tag{2.4}$$

where $\eta\colon \mathbb{R}^3 \to (0,\infty)$ is a smooth approximation of the point (delta) mass, with scale $\varepsilon_i > 0$ in the $x_i$ direction,

$$\eta(x) := \prod_{i=1}^{3}\frac{e^{-|x_i|^2/(2\varepsilon_i^2)}}{(2\pi\varepsilon_i^2)^{1/2}}. \tag{2.5}$$

Smooth averages have been used in molecular dynamics for fluid dynamics, cf. [18] and for the vortex blob method and the smoothed particle hydrodynamics approximation of moving particles in fluid dynamics, cf. [29], [2]. Sections 3–4 present a molecular dynamics model for the potential energy (2.4) and Section 5.4 formulates a kinetic Monte Carlo model.

**Question 2.1.** How accurate is it to say that the (macroscopic) latent heat is equal to a jump in $V$?

## 3. An atomistic Brownian dynamics model

The standard method to simulate molecular dynamics is to write Newton's laws for the particles, cf. [10], [32]. We will instead use Brownian dynamics with the Ito differential equations

$$dX_i^t = -\partial_{X_i} V(X^t)dt + \sqrt{2\gamma}\, dW_i^t \tag{3.1}$$

where $W_i$ are independent Brownian motions and the notation $X_i^t := X_i(t)$ is the position of the $i$'th particle at time $t$. This equation, called the Smoluchowski equation, is the zero relaxation time limit (i.e. $\tau \to 0+$) of Langevin's equation ( cf. [25], [30], [32], [21])

$$\begin{aligned} d\hat{X}_i^s &= p_i/\mu ds \\ dp_i^s &= -\partial_{X_i} V(\hat{X}^s)ds - \frac{p_i^s}{\tau}ds + \sqrt{\frac{2\gamma\mu}{\tau}}\, d\hat{W}_i^s, \end{aligned} \tag{3.2}$$

in the faster time scale $s = \mu t/\tau$, where $\mu$ is the mass and $\hat{W}_i$ are independent Brownian motions. The zero relaxation time limit is explained more in Remark 3.2. The simplified Brownian dynamics has the same invariant measure with density proportional to $e^{-V(X)/\gamma}$ as in Monte-Carlo molecular dynamics simulations of equilibrium problems with $\gamma = k_B T$, where $T$ is the absolute temperature and $k_B$ is the Boltzmann constant. In this sense, the parameter $\gamma/k_B$ in the Brownian dynamics is the local temperature $T$. In contrast to the standard Monte-Carlo method, the model (3.1) includes the time variable. Our microscopic model of a phase change is then the Brownian dynamics model (3.1) for the phase-field (latent heat) variable $m$ in (2.4) coupled to the macroscopic energy equation (2.1). The Brownian dynamics uses $\gamma := k_B T$, where the temperature varies on the macroscopic scale, due to the energy equation, so that $T$ is almost constant on the microscopic scale of a molecular dynamics simulation and makes its Gibbs equilibrium density proportional to $e^{-V(X)/(k_B T(x))}$ reasonable.

We have two reasons to use Brownian dynamics instead of standard deterministic Newton dynamics ($\tau = \infty$ in (3.2)): the most important reason is to have a formulation that separates the noise and the mean drift, which is a much harder issue in deterministic many particle dynamics, in fact so far the only derivation of the Euler equations of conservation laws from particle dynamics use a weak noise perturbation of a Hamiltonian system in [31]; and the second reason is to try to simulate molecular dynamics longer time.

**Question 3.1.** Is Brownian dynamics a reasonable alternative to standard molecular dynamics here?

**Remark 3.2** (Smoluchowski limit of the Langevin equation). The Smoluchowski high friction limit of the Langevin equation has been computed with different methods using strong [30] and weak convergence [25]. Strong convergence has the drawback to

yield error estimates of order $e^{Kt}\tau$, due to a Gronwall estimate and Lipschitz bound $K$ of the forces; in contrast, error estimates of probabilities using weak convergence can show good accuracy for long time. The proof that the Langevin solution $\hat{X}_{\mu t/\tau}$ converges weakly (i.e. in law) to the Smoluchowski solution $X_t$ as $\tau \to 0+$ in [25], [28] uses a Chapman–Enskog expansion of the Kolmogorov backward equation, for the Langevin dynamics in the diffusion time scale $t$, combined with a general convergence result for such diffusion processes in [26]. Dissipative particle dynamics [15] has dissipation-fluctuation perturbations of a Hamiltonian system where the momentum is conserved, in contrast to the analogous Langevin dynamics. The work [25] also shows that a Smoluchowski type limit seems more subtly for dissipative particle dynamics.

## 4. Coarse-grained phase-field dynamics

We want to determine a mean drift function $a(\bar{m})$ and a diffusion function $b(\bar{m})$ so that the coarse-grained approximation $\bar{m}^t$, solving the coarse-grained equation

$$d\bar{m}^t = a(\bar{m}^t)dt + \sum_{k=1}^{M} b_k(\bar{m}^t)d\tilde{W}_k^t,$$

is an optimal approximation to the phase field $m(X^t, \cdot)$ defined in (2.4), where $X^t$ solves the Brownian dynamics (3.1). Here $\tilde{W}_k$, $k = 1, \ldots, M$ are all independent Brownian motions, also independent of all $W_i$. For this purpose we seek to minimize the error of the expected value at any time $\mathcal{T}$

$$E[g(m(X^{\mathcal{T}}, \cdot))] - E[g(\bar{m}^{\mathcal{T}})]$$

for any given function $g$ with the same initial value $\bar{m}^0 = m(X^0, \cdot)$. Here the expected value of a stochastic variable $w$, with set of outcomes $\Omega$ and probability measure $P$, is defined by

$$E[w] := \int_{\Omega} w \, dP.$$

The first idea, in Section 4.1, is that Ito's formula and the Brownian dynamics (3.1) determine functions $\alpha$ and $\beta$, depending on the microscopic state $X$, so that

$$dm(X^t, \cdot) = \alpha(X^t)dt + \sum_{j=1}^{N} \beta_j(X^t)dW_j^t. \tag{4.1}$$

The next step, in Section 4.2, is to estimate the error, using the Kolmogorov equations for $\bar{m}$ and (4.1) similar to [35], [24], which leads to

$$E[g(m(X^{\mathcal{T}}, \cdot)) - g(\bar{m}^{\mathcal{T}})] = E\left[\int_0^{\mathcal{T}} \langle \bar{u}', \alpha - a\rangle + \langle \bar{u}'', \sum_{j=1}^{N} \beta_j \otimes \beta_j - \sum_{k=1}^{M} b_k \otimes b_k\rangle \, dt\right],$$

where $\langle \bar{u}', \cdot \rangle$ is the $L^2(\mathbb{R})$ scalar product corresponding to the variable $x$ with $\bar{u}'$, which is the Gateaux derivative (i.e. functional derivative) of the functional $E[g(\bar{m}^{\mathcal{T}}) \,|\, \bar{m}^t = n]$ with respect to $n$; and similarly $\langle \bar{u}'', \cdot \rangle$ is the $L^2(\mathbb{R} \times \mathbb{R})$ scalar product with the second Gateaux derivative $\bar{u}''$ of the functional $E[g(\bar{m}^{\mathcal{T}}) \,|\, \bar{m}^t = n]$ with respect to $n$. The notation $b_k \otimes b_k(x, x') := b_k(x) b_k(x')$ is the tensor product.

The final step, in Section 4.3, is to use molecular dynamics simulations for a planar two phase problem and take averages in cross sections parallel to the interface, where $\bar{u}', \bar{u}'', a, \sum_k b_k \otimes b_k$ are constant, to evaluate approximations to the functions $a$ and $\sum_k b_k \otimes b_k$ by

$$a = \frac{1}{\mathcal{T}} E\left[ \int_0^{\mathcal{T}} \alpha \, dt \right],$$

$$\sum_k b_k \otimes b_k = \frac{1}{\mathcal{T}} E\left[ \int_0^{\mathcal{T}} \sum_{j=1}^N \beta_j \otimes \beta_j \, dt \right].$$

**4.1. The Ito formula for the phase-field.** The Ito formula (cf. [13]) implies

$$dm(X^t, x) = \underbrace{\sum_{j=1}^N (-\partial_{X_j} m \, \partial_{X_j} V + \gamma \partial_{X_j X_j} m)}_{\alpha(X^t)} \, dt + \sum_{j=1}^N \underbrace{\sqrt{2\gamma} \partial_{X_j} m}_{\beta_j(X^t)} \, dW_j. \tag{4.2}$$

The definition in (2.4),

$$m(X^t, x) = \sum_i m_i(X) \eta(x - X_i^t),$$

yields

$$\partial_{X_j} m = \sum_i \partial_{X_j} m_i \eta(x - X_i) + m_j \partial_{X_j} \eta(x - X_j).$$

In (4.2) we will use (2.5) to evaluate the last derivative as

$$\partial_{X_j} \eta(x - X_j) = -\partial_x \eta(x - X_j) \quad \text{in } dt \text{ terms},$$

$$\partial_{X_j} \eta(x - X_j) = -\eta(x - X_j) \left( \frac{(x - X_j)_1}{\varepsilon_1^2}, \frac{(x - X_j)_2}{\varepsilon_2^2}, \frac{(x - X_j)_3}{\varepsilon_3^2} \right) \quad \text{in } dW_j \text{ terms},$$

in order to avoid spatial derivatives on the diffusion coefficient, while including them in the drift. Since

$$m_i = \frac{1}{2} \sum_{k \neq i} \Phi(X_i - X_k)$$

and

$$V(X) = \frac{1}{2} \sum_i \sum_{j \neq i} \Phi(X_i - X_j)$$

there holds

$$\partial_{X_j} m_i = \frac{1}{2}\sum_{k\neq i} \Phi'(X_i - X_k)\delta_{ij} - \frac{1}{2}\Phi'(X_i - X_j)(1-\delta_{ij}),$$
$$\partial_{X_j} V(X) = \sum_{k\neq j} \Phi'(X_j - X_k),$$

where

$$\delta_{ij} := \begin{cases} 1 & i = j, \\ 0 & i \neq j \end{cases}$$

is the Kronecker symbol. The second derivatives are

$$\partial_{X_j X_j} m = \sum_i \partial_{X_j X_j} m_i \eta(x - X_i) - 2\partial_{X_j} m_j \partial_x \eta(x - X_j) + m_j \partial_{xx}\eta(x - X_j),$$

with

$$\partial_{X_j X_j} m_i = \frac{1}{2}\sum_{k\neq i} \Phi''(X_i - X_k)\delta_{ij} + \frac{1}{2}\Phi''(X_i - X_j)(1-\delta_{ij})$$

and all terms in (4.2) are now expressed in terms of $\Phi$, its gradient $\Phi'$ and Hessian $\Phi''$. We note that the drift, $\alpha$, has the form

$$\partial_x\Big(\sum_{i=1}^N n_{2i}(X^t)\eta(x - X_i^t)\Big) + \sum_{i=1}^N n_{1i}(X^t)\eta(x - X_i^t)$$

of conservative and non conservative reaction terms. Similarly the diffusion, $\beta_j$, takes the form

$$\sum_{i=1}^N n_{3i}(X^t)\eta(x - X_i^t)(x - X_i^t).$$

**4.2. The error representation.** The conditioned expected value

$$\bar{u}(n,t) := E[g(\bar{m}^{\mathcal{T}}) \,|\, \bar{m}^t = n] \tag{4.3}$$

satisfies the Kolmogorov equation (cf. [35], [24])

$$\begin{aligned} &\partial_t u + \langle \bar{u}', a\rangle + \Big\langle \bar{u}'', \sum_{k=1}^M b_k \otimes b_k \Big\rangle = 0 \\ &\bar{u}(\cdot, \mathcal{T}) = g \end{aligned} \tag{4.4}$$

Let $m^t := m(X^t, \cdot)$. The final condition in (4.4) and the definition (4.3) show that

$$E[g(m(X^{\mathcal{T}}, \cdot)) - g(\bar{m}^{\mathcal{T}})] = E[\bar{u}(m^{\mathcal{T}}, \mathcal{T})] - \bar{u}(m^0, 0) = E\Big[\int_0^{\mathcal{T}} d\bar{u}(m^t, t)\Big].$$

Use the Ito formula and (4.2) to evaluate $d\bar{u}(m^t, t)$ and Kolmogorov's equation (4.4) to replace $\partial_t \bar{u}$ in this right hand side to obtain the error representation

$$
\begin{aligned}
&E[g(m(X^{\mathcal{T}}, \cdot)) - g(\bar{m}^{\mathcal{T}})] \\
&\quad = E\Big[\int_0^{\mathcal{T}} \langle \bar{u}', \alpha \rangle + \langle \bar{u}'', \sum_{j=1}^{N} \beta_j \otimes \beta_j \rangle + \partial_t \bar{u}\, dt\Big] \\
&\quad = E\Big[\int_0^{\mathcal{T}} \langle \bar{u}', \alpha - a \rangle + \langle \bar{u}'', \sum_{j=1}^{N} \beta_j \otimes \beta_j - \sum_{k=1}^{M} b_k \otimes b_k \rangle\, dt\Big].
\end{aligned}
$$

**4.3. Computation of averages in cross sections.** The optimal choice of the function $a$ is to minimize $E\big[\int_0^{\mathcal{T}} \langle \bar{u}', \alpha - a\rangle\, dt\big]$, which seems hard to determine exactly since $\bar{u}'(m(X^t, \cdot), t)$ depends on $X^t$. However, the function $\bar{u}'(m(X^t, \cdot), t)$ depends only mildly on the coarse-grained $m(X^t, \cdot)$ and not directly on $X^t$. Therefore a reasonable approximation of this optimum is to think of an expansion of $\bar{u}'$ in $\alpha - a$ and determine $a$ by the leading order condition $E\big[\int_0^{\mathcal{T}} \alpha - a\, dt\big] = 0$, which means that the drift $\bar{a}(x) := a(\bar{m}(\cdot, x))$ is

$$
\bar{a}(x) = \frac{1}{\mathcal{T}} E\Big[\int_0^{\mathcal{T}} \alpha(x)\, dt\Big],
$$

and similarly for the diffusion matrix

$$
\bar{d}(x, x') = \frac{1}{\mathcal{T}} E\Big[\int_0^{\mathcal{T}} \sum_{j=1}^{N} \beta_j(x) \otimes \beta_j(x')\, dt\Big].
$$

We expect the spatial averages of the microscopic variables to vary on a much smaller scale in the $x_1$ direction normal to the phase front than in its orthogonal directions. Consequently we use an average function $\eta$ in (2.4) with higher resolution in the $x_1$ direction, so that $0 < \varepsilon_1 \ll \varepsilon_2 = \varepsilon_3$. In a microscopic simulation the molecular dynamics (3.1) has a small spatial volume, so that $\varepsilon_2$ is much larger than the size of the simulation box. Consequently we may first think of $\alpha$ and $\beta$ depending only on the $x_1$ coordinate.

In practice, the drift $\bar{a}$ and diffusion $\bar{d}$ can only be determined for a discrete set of points

$$
\{(x_1(1), x_2(1), x_3(1)), \ldots, (x_1(M/3), x_2(M/3)), x_3(M/3))\} =: \mathcal{X}_M
$$

and $\mathcal{X}_M \times \mathcal{X}_M$, respectively, related to the scales $\varepsilon_i$. The diffusion coefficient $\bar{b}$, as a function of $x$, can then be obtained from Choleski factorization of the $M \times M$ matrix $\bar{d}$

$$
\sum_{k=1}^{M} \bar{b}_k(x) \bar{b}_k(x') = \bar{d}(x, x').
$$

We expect that $x_1 \mapsto \mathcal{T}^{-1} E\big[\int_0^{\mathcal{T}} m^t\, dt\big]$ is monotone, for fixed $(x_2, x_3)$, so that its inverse function, denoted by $m^{-1}$, is well defined. Then the coarse-grained drift and the diffusion can be obtained as function of $\bar{m}$ by

$$a(\bar{m}) := \bar{a}(m^{-1}(\bar{m})),$$

and similarly for $b_j$.

**Question 4.1.** Will the computed $a$ and $b$ be reasonable?

**Question 4.2.** Can the phase-field method be coupled to the molecular dynamics method for improved localized resolution?

**Question 4.3.** Note that the approximation error $E[g(m(X^{\mathcal{T}}, \cdot)) - g(\bar{m}^{\mathcal{T}})]$ becomes proportional to the variances

$$E\bigg[\int_0^{\mathcal{T}} \langle \alpha - a, \alpha - a\rangle\, dt\bigg],$$

$$E\bigg[\int_0^{\mathcal{T}} \Big\langle \sum_{j=1}^{N} \beta_j \otimes \beta_j - \sum_{k=1}^{M} b_k \otimes b_k, \sum_{j=1}^{N} \beta_j \otimes \beta_j - \sum_{k=1}^{M} b_k \otimes b_k \Big\rangle dt\bigg].$$

The first variations $\partial \bar{u}'(m(X^t, \cdot), t)/\partial\alpha$ and $\partial \bar{u}''(m(X^t, \cdot), t)/\partial\beta_j$ determine the factors of proportionality. Can this be used to adaptively determine the resolution scale $\varepsilon$?

**Remark 4.4.** If we integrate the noise term over all $x_1$, i.e. take $\varepsilon_1$ very large, and let $g(m) = m^2$, then the error $E[g(m(X^{\mathcal{T}}, \cdot)) - g(\bar{m}^{\mathcal{T}})]$ we are studying is the usual fluctuation of energy $E[V^2 - E[V]^2]$ (proportional to the specific heat [21]), provided we set $\bar{m} = E[V]$.

## 5. An atomistic kinetic Monte Carlo method

Kinetic Monte Carlo methods can also be used to simulate solid-liquid phase changes on an atomistic level, cf. [14]. Here the reaction states and rates are given a priori, which makes it possible to simulate crystal growth on larger time scales than in molecular dynamics. The reaction rates and states can in principle be determined from a molecular dynamics simulations on smaller systems, cf. [37]; however often several reactions are involved making this a demanding modeling task. This section is a short version of [24] and derives stochastic hydrodynamical limits of the Ising model with long range interaction, which is the simplest model of this kind of an stochastic interacting particle system on a square lattice with two possible states in each lattice point, cf. [21].

Define a periodic lattice $\mathcal{L} := \gamma Z^d \cap [0, 1]^d$, with neighboring sites on distance $\gamma$, and consider spin configurations $\sigma : \mathcal{L} \times [0, \mathcal{T}] \to \{-1, 1\}$ defined on this lattice.

Introduce a stochastic spin system where the spin $\sigma_t(x)$, at site $x \in \mathcal{L}$ and time $t$, will flip to $-\sigma_t(x)$ with the rate $c(x, \sigma_t(\cdot))dt$, in the time interval $(t, t+dt)$, independent of possible flips at other sites, cf. [27]. Let $\sigma^x$ denote the configuration of spins after a flip at $x$ of state $\sigma$, i.e.

$$\sigma^x(y) = \begin{cases} \sigma(y) & y \neq x, \\ -\sigma(x) & y = x, \end{cases}$$

the probability density $P(\sigma, t)$ of finding the spin system in configuration $\sigma \in \{-1, 1\}^{\mathcal{L}}$ at time $t$ then solves the master equation

$$\frac{dP(\sigma, t)}{dt} = \sum_{x \in \mathcal{L}} (c(x, \sigma^x)P(\sigma^x, t) - c(x, \sigma)P(\sigma, t)), \tag{5.1}$$

where the gain term $\sum_x c(x, \sigma^x)P(\sigma^x, t)$ is the probability of jumping to state $\sigma$ at time $t$ and the loss term $\sum_x c(x, \sigma)P(\sigma, t)$ is the probability to leave state $\sigma$. Similar master equations are used for microscopic models of chemical reactions and phase transformations, cf. [36], [14], where lattice sites are occupied by different species of particles. For instance with two species the state space could be $\{0, 1\} \times \{0, 1\}$ instead of $\{-1, 1\}$ for the classical spin model above.

We want a spin system that has statistical mechanics relevance, which can achieved e.g. by choosing the rate function $c$ as follows. Consider the Hamiltonian

$$H(\sigma) = -\frac{1}{2} \sum_{x \in \mathcal{L}} \sum_{y \neq x} J(x - y)\sigma(x)\sigma(y) - \sum_{x \in \mathcal{L}} h(x)\sigma(x)$$

$$J = \gamma^d J_0, \qquad J_0(x) = 0 \quad \text{for } |x| \geq 1,$$

where the long range interaction potential, $J_0 \in \mathcal{C}^2(\mathbb{R}^d)$, is compactly supported and the function $h \in \mathcal{C}^2(\mathbb{R}^d)$ is a given external field. Define the Glauber Markov process on $\{-1, 1\}^{\mathcal{L}}$ with generator

$$\frac{d}{dt} E[f(\sigma_t)|\sigma] = Lf(\sigma) = \sum_{x \in \mathcal{L}} c(x, \sigma)(f(\sigma^x) - f(\sigma)) \tag{5.2}$$

for $f : \{-1, 1\}^{\mathcal{L}} \to \mathbb{R}$ and the flip rate

$$\begin{aligned} c(x, \sigma) &= \frac{e^{-\beta U(x)\sigma(x)}}{e^{-\beta U(x)} + e^{\beta U(x)}} = \frac{1}{2}\big(1 - \sigma(x)\tanh(\beta U(x))\big), \\ U(x) &= h(x) + \sum_{y \neq x} J(x - y)\sigma(y) =: h(x) + J * \sigma(x) - J(0)\sigma(x), \end{aligned} \tag{5.3}$$

where $\beta > 0$ is the inverse temperature. This flip rate has built in invariance of the Gibbs density, $e^{-\beta H(\sigma)} / \sum_\sigma e^{-\beta H(\sigma)}$, since it satisfies the detailed balance

$$c(x, \sigma)e^{-\beta H(\sigma)} = c(x, \sigma^x)e^{-\beta H(\sigma^x)},$$

which implies that this Gibbs density is a time independent (invariant) solution to (5.1). Having this invariant Gibbs measure implies that the model has statistical mechanics relevance, see [12], [4], [5], [6], [11]. For example in a neighborhood of $x \in \mathcal{L}$, where $h$ and $J * (1, \ldots, 1)$ are positive, the construction of the flip rate $c$ makes the system favor phases with spins mostly equal to 1 as compared to phases with spins mostly equal to $-1$.

We will study localized projection averages of $\sigma$ on scale $\varepsilon$. In particular we will find approximations to expected values of such averages. The error analysis uses consistency with the backward equation

$$\partial_t \tilde{u} + L\tilde{u} = 0 \quad \text{for } t < \mathcal{T}, \quad \tilde{u}(\cdot, \mathcal{T}) = g$$

corresponding to the master equation (5.1) for expected values

$$\tilde{u}(\xi, t) := E[g(\sigma_{\mathcal{T}})|\, \sigma_t = \xi].$$

**5.1. A coarse-grained kinetic Monte Carlo method.** Define the coarse periodic lattice $\bar{\mathcal{L}} := q\gamma Z^d \cap [0, 1]^d$ with neighboring sites on distance $q\gamma =: \varepsilon$, where $q$ is an even positive integer and $q^d$ is the number of fine sites projected to a coarse site: the lattice points $y \in \bar{\mathcal{L}}$ define the coarse cells

$$C_y = \{x \in \mathcal{L} : -q\gamma/2 \le x_i - y_i < q\gamma/2\},$$

of $q^d$ neighboring points in the fine lattice and the averaging operator

$$A_\varepsilon(z, x) = \begin{cases} 1/q^d & \text{if } x \text{ and } z \text{ are in the same coarse cell } C_y, \\ 0 & \text{if } x \text{ and } z \text{ are in different coarse cells.} \end{cases}$$

We will study the behavior of the localized projection averages

$$\overline{X}(z) := \sum_{x \in \mathcal{L}} A_\varepsilon(z, x)\sigma(x), \tag{5.4}$$

for $z \in \mathcal{L}$. The coarse-grained average $\overline{X}$ can be interpreted as a function on the coarse lattice since the restriction of $\overline{X}$ to each coarse cell $C_z$ is constant, i.e. $\overline{X} = \sum_{x \in C_\cdot} \sigma(x)/q^d$.

The work [23] derives a coarse-grained kinetic Monte Carlo equation approximating the average $\overline{X}$. The next section shows as in [24] that the average spin, $\overline{X}$, can be approximated by the solution, $X \colon \bar{\mathcal{L}} \times [0, \mathcal{T}] \times \Omega \to \mathbb{R}$, to the Ito stochastic differential equation

$$dX_t(x) = a(X_t)(x)dt + b(X_t)(x)dW^x, \quad X_0 = \overline{X}_0, \tag{5.5}$$

with the drift, $a\colon \mathbb{R}^{\bar{\mathcal{L}}} \to \mathbb{R}^{\bar{\mathcal{L}}}$, and diffusion, $b\colon \mathbb{R}^{\bar{\mathcal{L}}} \to \mathbb{R}^{\bar{\mathcal{L}}}$, coefficients given by

$$\begin{aligned} a(X) &= -X + \tanh(\beta(J * X + h - J(0)X)),\\ b(X)(x) &= \left(\frac{\gamma}{\varepsilon}\right)^{d/2} \sqrt{|1 - X\tanh(\beta(J * X + h - J(0)X))(x)|}\ \eta(X(x)),\\ \eta(r) &= \begin{cases} 1 & \text{for } x \in [-1, 1],\\ 0 & \text{for } x \in (-\infty, -\hat{r}) \cup (\hat{r}, \infty),\end{cases}\\ \hat{r} &:= \min(1 + e^{-2\beta(2|J|_{\ell^1} + \|h\|_{L^\infty})}, 3/2) \end{aligned} \tag{5.6}$$

and a Wiener process $W\colon \bar{\mathcal{L}} \times [0, \mathcal{T}] \times \Omega \to \mathbb{R}$ on a probability space $(\Omega, P, \{\mathcal{F}_t\}_{t=0}^{\mathcal{T}})$, with the set of outcomes $\Omega$, probability measure $P$ and sigma algebra $\mathcal{F}_t$ of events up to time $t$. Here $W^x$ are independent one dimensional standard Brownian motions for $x \in \bar{\mathcal{L}}$, so that formally

$$\begin{aligned} E[dW_t^x] &= 0,\\ E[dW_s^x dW_t^y] &= 0 \quad \text{for } s \neq t,\\ E[dW_t^x dW_t^y] &= 0 \quad \text{for } x \neq y, \text{ and}\\ E[dW_t^x dW_t^x] &= dt. \end{aligned}$$

The $\mathcal{C}^\infty$ cut-off function $\eta\colon \mathbb{R} \to [0, 1]$, with compact support, is introduced to handle the complication that $|X(x)|$ may be larger than 1, although $|\overline{X}(x)|$ is not, so that $1 - X\tanh(\beta(J * X + h - J(0)X))(x)$ may be close to zero causing large values on derivatives of

$$\sqrt{|1 - X\tanh(\beta(J * X + h - J(0)X))(x)|}\,,$$

note that we have $|\overline{X}(x)| \leq 1$ and consequently the cut-off $\eta$ improves the approximation by switching off the noise before $1 - X\tanh(\beta(J * X + h - J(0)X))(x)$ becomes zero making $b$ a $\mathcal{C}^\infty$ function.

The approximation uses that the high dimensional value function $u\colon \mathbb{R}^{\bar{\mathcal{L}}} \times [0, \mathcal{T}] \to \mathbb{R}$ defined by

$$u(\xi, t) = E[g(X_{\mathcal{T}})|X_t = \xi]$$

solves a corresponding Kolmogorov backward equation, where the drift and diffusion coefficients in (5.6) are chosen to minimize the error $E[g(\overline{X}_{\mathcal{T}})] - E[g(X_{\mathcal{T}})]$. To define the Kolmogorov backward equation introduce the weighted scalar products

$$\begin{aligned} w \cdot v &:= \sum_{y \in \bar{\mathcal{L}}} w_y v_y \varepsilon^d && \text{for } w, v \in \ell^2(\bar{\mathcal{L}}),\\ w \cdot v &:= \sum_{x,y \in \bar{\mathcal{L}}} w_{xy} v_{xy} \varepsilon^{2d} && \text{for } w, v \in \ell^2(\bar{\mathcal{L}}^2),\\ w \cdot v &:= \sum_{x,y,z \in \bar{\mathcal{L}}} w_{xyz} v_{xyz} \varepsilon^{3d} && \text{for } w, v \in \ell^2(\bar{\mathcal{L}}^3). \end{aligned}$$

Then $u$ satisfies the Kolmogorov backward equation

$$\partial_t u + a \cdot u' + D \cdot u'' = 0 \quad \text{for } t < \mathcal{T},$$
$$u(\cdot, \mathcal{T}) = g,$$

where

$$D_{xy} = \begin{cases} (1 - X \tanh(\beta(J * X + h))(x))\eta^2(X(x)) & y = x, \\ 0 & y \neq x, \end{cases}$$

and $u'(\xi, t) = \partial_\xi u(\xi, t)$ and $u''(\xi, t)$ are the first and second order Gateaux derivatives of $u$ in $\ell^2(\bar{\mathcal{L}})$ and $\ell^2(\bar{\mathcal{L}}^2)$, respectively.

**5.2. Stochastic hydrodynamical limit of the particle system.** The main result in [24] is

**Theorem 5.1.** *The average spin, $\overline{X}$, can be approximated by the solution, $X$, to the Ito stochastic differential equation* (5.5) *with error*

$$E[g(\overline{X}_{\mathcal{T}})] - E[g(X_{\mathcal{T}})] = \mathcal{O}\big(\mathcal{T}\varepsilon + \mathcal{T}(\gamma/\varepsilon)^{2d}\big) \tag{5.7}$$

*provided that the Gateaux derivatives $u'(\overline{X}_t, t)$, $u''(\overline{X}_t, t)$ and $u'''(\overline{X}_t, t)$ on the path $\overline{X}$ are bounded in the weighted norms $\ell^1(\bar{\mathcal{L}}^i)$ up to time $\mathcal{T}$.*

Note that $a = 0$ gives $\mathcal{O}(1)$ error, while $b = 0$ gives $\mathcal{O}((\gamma/\varepsilon)^d)$ error so that $b$ defined by (5.6) is justified for $\gamma \ll \varepsilon \ll \gamma^{2d/(2d+1)}$, with $\mathcal{T}$ fixed.

The stochastic differential equation (5.5) has $\mathcal{C}^\infty$ coefficients, where perturbations of solutions may grove exponentially in time. The work [24] verifies that mean square estimates of $X$ and its variations up to order three give bounds on the weighted $\ell^1$-norm of the derivatives of $u$ that depend exponentially on time, i.e. $e^{C\mathcal{T}}$.

*Proof of the theorem.* The definitions of $u$, the generator (5.2) and the average (5.4) imply

$$\begin{aligned}
&E[g(\overline{X}_{\mathcal{T}})] - E[g(X_{\mathcal{T}})] = E[u(\overline{X}_{\mathcal{T}}, \mathcal{T})] - E[u(X_0, 0)] \\
&= E\bigg[\int_0^{\mathcal{T}} du(\overline{X}_t, t)\bigg] \\
&= \int_0^{\mathcal{T}} E[Lu + \partial_t u]\, dt = \int_0^{\mathcal{T}} E\big[E[Lu - a \cdot u' - D \cdot u''|\overline{X}_t]\big]\, dt \\
&= \int_0^{\mathcal{T}} E\Big[E\Big[\sum_{x \in \mathcal{L}} c(x, \sigma)\big(u(\overline{X}(\sigma^x)) - u(\overline{X}(\sigma))\big) - a \cdot u' - D \cdot u''|\overline{X}_t\Big]\Big]\, dt \\
&= \int_0^{\mathcal{T}} E\Big[E\Big[\sum_{x \in \mathcal{L}} c(x, \sigma)\big(u(\overline{X}(\sigma) - 2A_\varepsilon(x, \cdot)\sigma(x)) - u(\overline{X}(\sigma))\big) \\
&\qquad\qquad - a \cdot u' - D \cdot u''|\overline{X}_t\Big]\Big]\, dt.
\end{aligned} \tag{5.8}$$

The first step to estimate this error is to write the differences in $u$ in terms of its Gateaux derivatives by Taylor expansion, for some $s \in [0,1]$,

$$\begin{aligned} &u(\overline{X}(\sigma) - 2A_\varepsilon(x,\cdot)\sigma(x)) - u(\overline{X}(\sigma)) \\ &\quad = -2u'(\overline{X}) \cdot A_\varepsilon(x,\cdot)\sigma(x) \\ &\qquad + 2u''(\overline{X}) \cdot A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)\sigma^2(x) \\ &\qquad - \frac{4}{3}u'''(\overline{X} - 2sA_\varepsilon(x,\cdot)\sigma(x)) \cdot A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)\sigma^3(x), \end{aligned} \tag{5.9}$$

so that the error representation (5.8) becomes

$$\begin{aligned} &E[g(\overline{X}_{\mathcal{T}})] - E[g(X_{\mathcal{T}})] \\ &= \int_0^{\mathcal{T}} E\Big[E\Big[\sum_{x\in\mathcal{L}} \big(u'(\overline{X}) \cdot (-2c(x,\sigma)A_\varepsilon(x,\cdot)\sigma(x) - a) \\ &\quad + u''(\overline{X}) \cdot (2c(x,\sigma)A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)\sigma^2(x) - D) \\ &\quad - \frac{4}{3}u'''(\overline{X} - 2sA_\varepsilon(x,\cdot)\sigma(x)) \cdot c(x,\sigma)A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)\sigma^3(x)\big) \mid \overline{X}_t\Big]\Big]\, dt. \end{aligned} \tag{5.10}$$

The next step is to determine the optimal $a$ and $b$ which minimize the error (5.10). For this purpose we shall in the flipping rate approximate the coupling $J * \sigma$ and $J(0)\sigma = \mathcal{O}(\gamma^d)$ with $J * \overline{X}$ and $J(0)\overline{X}$, using the long range $\mathcal{O}(1)$ interaction distance of $J$. The definition of the average (5.4) implies

$$J * \overline{X} = \sum_{z,y\in\mathcal{L}} J(\cdot - y)A_\varepsilon(y,z)\sigma(z)$$

and consequently the coupling has the uniform error estimate

$$\|J*\sigma - J*\overline{X}\|_{\ell^\infty} \le \Big\|J(\cdot - z) - \sum_{y\in\mathcal{L}} J(\cdot - y)A_\varepsilon(y,z)\Big\|_{\ell^1} \|\sigma\|_{\ell^\infty} = \mathcal{O}(\varepsilon). \tag{5.11}$$

This error estimate, the flip rate (5.3) and $J(0) = \mathcal{O}(\gamma^d)$ imply

$$\begin{aligned} &-\sum_{x\in\mathcal{L}} 2c(x,\sigma)A_\varepsilon(x,\cdot)\sigma(x) \\ &\qquad = -\overline{X} + A_\varepsilon \cdot \tanh(\beta(J*\sigma + h - J(0)\sigma)) \\ &\qquad = -\overline{X} + \tanh(\beta(J*\overline{X} + h - J(0)\overline{X})) + \mathcal{O}(\varepsilon + \gamma^d), \end{aligned} \tag{5.12}$$

and

$$\begin{aligned} &\sum_{x\in\mathcal{L}} 2c(x,\sigma)A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot)\sigma^2(x) \\ &\qquad = \Big(\frac{\gamma}{\varepsilon}\Big)^d \big[1 - \overline{X}\tanh(\beta(J*\overline{X} + h - J(0)\overline{X}))\big] \\ &\qquad\quad + \mathcal{O}\big((\gamma/\varepsilon)^{2d} + \varepsilon + \gamma^{2d}\big). \end{aligned} \tag{5.13}$$

We have

$$\begin{aligned} \Big\| \sum_x A_\varepsilon(x,\cdot)A_\varepsilon(x,\cdot) \Big\|_{\ell^\infty} &= (\gamma/\varepsilon)^d, \\ \Big\| \sum_x A_\varepsilon(x-\cdot)A_\varepsilon(x-\cdot)A_\varepsilon(x-\cdot) \Big\|_{\ell^\infty} &= (\gamma/\varepsilon)^{2d}, \end{aligned} \tag{5.14}$$

which together with the expansions (5.10), (5.12) and (5.13) proves the theorem. □

We also have

**Lemma 5.2.** *Suppose that the Gateaux derivatives $u'(\overline{X}_t, t)$ and $u''(\overline{X}_t, t)$ on the path $\overline{X}$ are bounded in the weighted norms $\ell^1(\bar{\mathcal{L}}^i)$ up to time $\mathcal{T}$ and that the initial spin $\sigma_0$ has expected value $m$, where $\sigma_0(x) - m_x$ are i.i.d. with bounded variance and second order difference quotients $|d^2m/dx^2| = \mathcal{O}(1)$. Then the deterministic mean field solution, $\hat{X}\colon \mathbb{R}^{\bar{\mathcal{L}}} \times [0, \mathcal{T}] \to \mathbb{R}$,*

$$d\hat{X}/dt = -\hat{X} + \tanh(\beta(J * \hat{X} + h - J(0)\hat{X})), \quad \hat{X}_0 = E[\overline{X}_0],$$

*depends on $\varepsilon$ only through the initial data and satisfies*

$$E[g(\overline{X}_{\mathcal{T}})] - E[g(\hat{X}_{\mathcal{T}})] = \mathcal{O}(\varepsilon + (\gamma/\varepsilon)^d)$$

*provided the drift $a$ is defined by* (5.6).

*Proof.* Think of $\hat{X}$ as an $X$ with $b = 0$ and apply the corresponding expansion (5.8), (5.9) and (5.14). Then it remains to verify that the initial data satisfy

$$E[u(\overline{X}_0, 0) - u(\hat{X}_0, 0)] = \mathcal{O}\big((\gamma/\varepsilon)^d\big),$$

but this is a direct consequence of the central limit theorem and the initial $\sigma_0 - E[\sigma_0]$ being i.i.d. with bounded variance. □

**5.3. Alternative invariant measure diffusion for mean exit times.** Not all expected values $E[g(\overline{X}_{\mathcal{T}})]$ can be approximated using the stochastic differential equation (5.5) with Einstein diffusion, due to the required bounds on the derivatives of $u$; such an example is to determine the expected first exit time $\tau(Y) = \inf\{t : Y_t \notin A\}$ from a neighborhood $A$ of an equilibrium point $y' \in A$, where $a(y') = 0$ and $Y_0 \in A$. Then the expected exit time is exponentially large, i.e.

$$\begin{aligned} &\lim_{\gamma/\varepsilon \to 0+} \left(\tfrac{\gamma}{\varepsilon}\right)^d \log E[\tau(\overline{X})] \text{ and } \lim_{\gamma/\varepsilon \to 0+} \left(\tfrac{\gamma}{\varepsilon}\right)^d \log E[\tau(X)] \\ &\text{are both strictly positive.} \end{aligned} \tag{5.15}$$

These expected values are related to transition rates $k$ and $E[\tau] = 1/k$ in simple cases, see [17], [9]. Hanggi et al. [16] have proposed a remedy by approximating the master equation by a different stochastic differential equation with the same asymptotic drift

but a modified diffusion, to leading order, chosen so that the SDE invariant density $Z^{-1}e^{-U/(\gamma/\varepsilon)^d}$ is asymptotically the same as for the master equation. One perspective on the two different SDEs with Einstein diffusion or invariant measure diffusion is that the two limits, coarse-graining and time tending to infinity, do not commute. Because of (5.15) the theory of large deviations for rare events is relevant for exit times, cf. [9].

Let $\gamma_1 := \gamma/\varepsilon$. Consider an SDE

$$dX_t(x) = (a(X_t) + \gamma_1^d c(X_t))(x)dt + \gamma_1^{d/2}\tilde{b}(X_t)(x)dW_t^x,$$

with the generator

$$Lf = (a + \gamma_1^d c)\cdot f' + \gamma_1^d \tilde{D}\cdot f'', \quad D_{ij} = \tilde{b}_i\tilde{b}_j\delta_{ij};$$

the idea in [16] is to find $c$ and $D$ so that the corresponding SDE asymptotically has the same invariant density $e^{-U/\gamma_1^d}/Z$ as the master equation. Hanggi et al. [16] determine the diagonal diffusion matrix $\tilde{D}$ and the small contribution to the drift $\gamma_1^d c$ by

$$\begin{aligned} \tilde{D}_{ii} &= -a_i/U_i', \\ c_i &= -\partial_{x_i}\tilde{D}_{ii}; \end{aligned} \tag{5.16}$$

note that since $a$ and $U$ have the same zeros, the constructed function $\tilde{D}_{ii}$ is positive in general. The equation (5.16) can be obtained by the WKB expansion

$$\begin{aligned} 0 \simeq L^* e^{-U/\gamma_1^d} = \big(&\gamma_1^{-d}(a_i U_i' + \tilde{D}_{ii}U_i'U_i') \\ &+ \gamma_1^0(\partial_i a_i + 2U_i'\partial_i\tilde{D}_{ii} + U_{ii}''\tilde{D}_{ii} + c_i U_i') \\ &+ \gamma_1^d(\partial_i c - \partial_{ii}\tilde{D}_{ii})\big)e^{-U/\gamma_1^d} \end{aligned}$$

together with the two leading order conditions that the terms of order $\gamma_1^{-d}$ and $\gamma_1^0$ vanish; here $L^*$ is the Chapman–Enskog operator adjoint to $L$. Consequently the choice (5.16) will in general generate an SDE with an invariant density $e^{-\tilde{U}/\gamma_1^d}/Z$, where $|\tilde{U} - U| = \mathcal{O}(\gamma_1^{2d})$.

Let us indicate why good approximation of the invariant measure implies that also the expected values, $E[\tau]$, for exit problems related to rare events with large deviations, are accurately computed: the work [9] shows that

$$\lim_{\gamma_1\to 0+} \gamma_1^d \log E[\tau(X)] = \inf_{y\in\partial A} U(y) - U(y'), \tag{5.17}$$

for one stable attracting equilibrium point $y' \in A$. The work [24] shows that the exit time (5.17) with SDE's and invariant measure diffusion is asymptotically the same as for the master equation for the 1D Curie–Weiss model:

$$\lim_{\gamma_1\to 0+} \gamma_1^d\big(\log E[\tau(X)] - \log E[\bar{\tau}(\overline{X})]\big) = 0\,, \tag{5.18}$$

where $E[\tau(X)]$ and $E[\bar{\tau}(\overline{X})]$ denote the mean exit time for the Hanggi SDE and the Curie–Weiss master equation, respectively. The Curie–Weiss model is a simple adsorption/desorption Ising model with constant interaction potential, cf. Section 5.4. The technique to establish this asymptotic agreement is to use logarithmic (Hopf–Cole) transformations of the two mean exit times, as functions of the initial location, which transforms the corresponding two linear Kolmogorov backward equations to two nonlinear Hamilton–Jacobi equations, cf. [8]. The two processes give rise to two different asymptotic Hamilton–Jacobi equations, however the key observation is that they have the same viscosity solution since they are both convex and have the same set of zeros.

**5.4. Dendrites with Einstein diffusion.** We see by Theorem 5.1 and Lemma 5.2 that the mean field differential equation solution is also an accurate approximation to the spin dynamics, provided the derivatives of the value function are bounded; this indicates that the stochastic differential equation (5.5) then only offers a small quantitative improvement. However, if the derivatives of the value function are large the mean field solution may give a qualitatively wrong answer, with $\mathcal{O}(1)$ error as $\gamma/\varepsilon \to 0+$, while the stochastic differential equation still yields an asymptotically correct limit; such an example is dendrite formation in phase transformations, cf. [22], [19], [3], [14].

Let us try to motivate why the noise in Theorem 5.1 seems applicable to dendrite formation. Dendrite dynamics can be formulated by the phase field method with an Allen–Cahn/Ginzburg–Landau equation coupled to a diffusion equation for the energy, as in (1.1), and by master equations coupled to the energy equation, cf. [14]. Mean field equations related to such a phase field system have been derived from a spin system coupled to a diffusion equation, see [7].

A master equation variant of the molecular dynamics model in Sections 2–4 is to let the coarse-grained potential energy be defined by

$$m(\sigma, z) := \sum_x \Big( \sum_{y \neq x} \frac{1}{2} J(x-y)\sigma(y) - h \Big) \sigma(x) A_\varepsilon(x, z),$$

where $A$ is the average in (5.4), and replace the Glauber dynamics with Arrhenius dynamics. That is, the microscopic dynamics is given by independent spins $\sigma(x) \in \{0, 1\}$, for each lattice point $x \in \mathcal{L}$, flipping with adsorption rate

$$c_a(x) = d_0\big(1 - \sigma(x)\big)$$

from states 0 to 1, and with desorption rate

$$c_d(x) = d_0\sigma(x) \exp\Big( -\frac{1}{k_B T} \Big( \sum_{y \neq x} J(x-y)\sigma(y) - h \Big) \Big)$$

from states 1 to 0, where $h$ is a surface binding energy or an external field and $d_0$ is a given rate, cf. [23]. Arrhenius dynamics also satisfies detailed balance with the same

Gibbs density

$$e^{(\sum_x \sum_{y\neq x} J(x-y)\sigma(x)\sigma(y)/2 - \sum_x h\sigma(x))/(k_B T)}$$

as for Glauber dynamics. The dynamics for the potential energy variable can then be coupled to the energy equation (2.1)

$$\partial_t(c_v T + m) = \operatorname{div}(k\nabla T)$$

by letting the temperature $T$ vary on the coarse-grained scale.

The dendrite grows with a positive non vanishing speed. Without noise in the model there is no side branching, while the side branching is present with added noise to the phase field model, cf. [3], or to the mean field model derived in [14]. This noise induced side branching is explained by the high sensitivity with respect to small perturbations at the dendrite tip, cf. [22]. Therefore the derivatives of an appropriate value function are large. Here the value function, $u$, could for instance measure the total dendrite surface at a fixed time. The inconsistent approximation of the mean field solution could by Lemma 5.2 be explained by having

$$(\gamma/\varepsilon)^d \|u''\|_{\ell^1} = \mathcal{O}(1). \tag{5.19}$$

The smallest scale in the problem is the dendrite tip radius $\rho$; with a bounded value function its derivatives could then be

$$\begin{aligned} \|u'\|_{\ell^1} &= \mathcal{O}(1/\rho), \\ \|u''\|_{\ell^1} &= \mathcal{O}(1/\rho^2), \\ \|u'''\|_{\ell^1} &= \mathcal{O}(1/\rho^3). \end{aligned}$$

Consequently (5.19) yields $(\gamma/\varepsilon)^{d/2} = \rho$, so that the noise error for the stochastic differential equation with the Einstein diffusion of Theorem 5.1 would be bounded by $(\gamma/\varepsilon)^{2d}\|u'''\|_{\ell^1} = \mathcal{O}((\gamma/\varepsilon)^{d/2})$, which tends to zero as $\gamma/\varepsilon \to 0+$. Therefore, this adsorption/desorption kinetic Monte Carlo model with long range interaction generates an approximating stochastic differential equation, which could be applicable also to coupling with the energy equation if the derivation remains valid with slowly varying temperature. An essential and maybe more difficult question is to find accurate kinetic Monte Carlo methods for real systems with dendrite dynamics, e.g. using ideas from the molecular dynamics coarse-graining in Sections 2–4.

## References

[1] Amberg, G., Semi sharp phase-field method for quantitative phase change simulations. *Phys. Rev. Lett.* **91** (2003), 265505–265509.

[2] Beale, J. T., Majda, A. J., Vortex methods. I. Convergence in three dimensions. *Math. Comp.* **39** (1982), 1–27.

[3] Boettinger, W. J., Warren, J. A., Beckermann, C., Karma, A., Phase field simulation of solidification. *Ann. Rev. Mater. Res.* **32** (2002), 163–194.

[4] De Masi, A., Orlandi, E., Presutti, E., Triolo, L., Glauber evolution with the Kac potentials. I. Mesoscopic and macroscopic limits, interface dynamics. *Nonlinearity* **7** (1994), 633–696.

[5] De Masi, A., Orlandi, E., Presutti, E., Triolo, L., Glauber evolution with Kac potentials. II. Fluctuations. *Nonlinearity* **9** (1996), 27–51.

[6] De Masi, A., Orlandi, E., Presutti, E., Triolo, L., Glauber evolution with Kac potentials. III. Spinodal decomposition. *Nonlinearity* **9** (1996), 53–114.

[7] Dirr, N., Luckhaus, S., Mesoscopic limit for non-isothermal phase transition. *Markov Processes and Related Fields* **7** (2001), 355–381.

[8] Fleming, W. H., Soner, H. M., *Controlled Markov Processes and Viscosity Solutions*. Appl. Math. (New York) 25, Springer-Verlag, New York 1993.

[9] Freidlin, M. I., Wentzell, A. D., *Random Perturbations of Dynamical Systems*. Grundlehren Math. Wiss. 260, Springer-Verlag, New York 1984.

[10] Frenkel, D., Smit, B., *Understanding Molecular Simulation*. Comput. Sci. Ser. 1, Academic Press, Orlando, FL, 2002.

[11] Giacomin, G., Lebowitz, J. L., Presutti, E., Deterministic and stochastic hydrodynamic equations arising from simple microscopic model systems. In *Stochastic partial differential equations: six perspectives*, Math. Surveys Monogr. 64, Amer. Math. Soc., Providence, RI, 1999, 107–152.

[12] Glauber, R. J., Time-dependent statistics of the Ising model. *J. Math. Phys.* **4** (1963), 294–307.

[13] Goodman, J., Moon, K.-S., Szepessy, A., Tempone, R., Zouraris, G., Stochastic and partial differential equations with adapted numerics. http://www.math.kth.se/~szepessy/sdepde.pdf

[14] Gouyet, J. F., Plapp, M., Dietrich, W., Maass, P., Description of far-from-equilibrium process by mean-field lattice gas models. *Adv. in Phys.* **52** (2003), 523–638.

[15] Groot, R. D., Warren, P. B., Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **107** (1997), 4423–4435.

[16] Hanggi, P., Grabert, H., Talkner, P., Thomas, H., Bistable systems: master equation versus Fokker-Planck modeling. *Phys. Rev. A* **3** (1984), 371–378.

[17] Hanggi, P., Talkner, P., Borkovec, M., Reaction-rate theory - 50 years after Kramers. *Rev. Modern Phys.* **62** (1990), 251–341.

[18] Hardy R. J., Formulas for determining local properties in molecular dynamics: shock waves. *J. Chem. Phys.* 76 (1982), 622-628.

[19] Hoyt, J. J., Asta, M., Karma, A., Atomistic and continuum modeling of dendritic solidification, *Materials Science Engineering R-Reports* **41** (2003), 121–163.

[20] Irving, J. H., Kirkwood, J. G., The statistical mechanics of transport processes: IV the equations of hydrodynamics. *J. Chem. Phys.* **18** (1950), 817–829.

[21] Kadanoff, L. P., *Statistical physics : statics, dynamics and renormalization*. World Scientific, Singapore 2000.

[22] Karma, A., Rappel, W. J., Phase-field model of dendritic side branching with thermal noise. *Phys Rev. E.* **60** (1999), 3614–3625.

[23] Katsoulakis, M. A., Majda, A. J., Vlachos, D. G., Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems. *J. Comput. Phys.* **186** (2003), 250–278.

[24] Katsoulakis, M., Szepessy, A., Stochastic hydrodynamical limits of particle systems. Preprint, 2005; http://www.nada.kth.se/∼szepessy/

[25] Kramer, P. R., Majda, A. J., Stochastic mode reduction for particle-based simulation methods for complex microfluid systems. *SIAM J. Appl. Math.* **64** (2003), 401–422.

[26] Kurtz, T. G., A limit theorem for perturbed operator semigroups with applications to random evolutions. *J. Funct. Anal.* **12** (1973), 55–67.

[27] Liggett, T. M., *Interacting particle systems*. Classics Math., Springer-Verlag, Berlin 2005.

[28] Majda, A. J., Timofeyev, I., Vanden Eijnden, E., A mathematical framework for stochastic climate models. *Comm. Pure Appl. Math.* **54** (2001), 891–974.

[29] Mas-Gallic, S., Raviart, P.-A., A particle method for first-order symmetric systems. *Numer. Math.* **51** (1987), 323–352.

[30] Nelson, E., *Dynamical Theories of Brownian Motion*. Princeton University Press, Princeton, NJ, 1967.

[31] Olla, S., Varadhan, S. R. S., Yau, H.-T., Hydrodynamical limit for a Hamiltonian system with weak noise. *Comm. Math. Phys.* **155** (1993), 523–560.

[32] Schlick, T., *Molecular modeling and simulation*. Interdiscip. Appl. Math. 21, Springer-Verlag, New York 2002.

[33] Shardlow, T., Splitting for dissipative particle dynamics. *SIAM J. Sci. Comput.* **24** (2003), 1267–1282.

[34] Soner, H. M., Convergence of the phase-field equations to the Mullins-Sekerka Problem with kinetic undercooling. *Arch. Rat. Mech. Anal.* **31** (1995), 139–197.

[35] Szepessy, A., Tempone, R., Zouraris, G., Adaptive weak approximation of stochastic differential equations. *Comm. Pure Appl. Math.* **54** (2001), 1169-1214.

[36] van Kampen, N. G., *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam 1981.

[37] Yip, S. (ed.), *Handbook of Materials Modeling*. Springer-Verlag, Berlin 2005.

Institutionen för Matematik, Kungl. Tekniska Högskolan, 10044 Stockholm, Sweden
E-mail: szepessy@kth.se

# Competitions and mathematics education

Petar S. Kenderov*

**Abstract.** Mathematics competitions, together with the people and organizations engaged with them, form an immense and vibrant global network today. This network has many roles. Competitions help identify students with higher abilities in mathematics. They motivate these students to develop their talents and to seek professional realization in science. Competitions have positive impact on education and on educational institutions. Last but not least, a significant part of the classical mathematical heritage known as "Elementary Mathematics" is preserved, kept alive and developed through the network of competitions and competition-related activities. Nevertheless, competitions need to evolve in order to meet the demands of the new century.

These and many other items are outlined and discussed in the paper.

**Mathematics Subject Classification (2000).** Primary 97U40; Secondary 97C60.

**Keywords.** Mathematics competitions, olympiads, higher ability students.

## 1. Introduction

Competition is essential and intrinsic to life. Every day, living things in nature and economic subjects in society compete for resources, for better living conditions, and for higher efficiency. The desire to compete in overcoming a challenge is deeply rooted in human nature and has been employed for centuries to help people sharpen their skills and improve their performance in various activities.

Competitions, however hotly debated, praised, or condemned, remain central and inherent in education. Both the traditional marking (grading) of students in school and the more innovative measuring of their basic scholastic abilities (implemented by methods such as PISA, TIMSS, or SAT) inevitably create, directly or indirectly, competition among students, among teachers, among schools, and even among whole countries. Heated debates aside, few would deny the positive influence such competitions bring to the process of teaching and learning, and to the overall performance of the educational system.

The interaction between competition and education is more complex, however. It is not only that competitions enhance education. Education itself can be viewed as

*The author is grateful to the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, and to his colleagues for their support of (and the involvement with) the Bulgarian system of competitions in mathematics, informatics, and mathematical linguistics. Special thanks to E. Belogay, participant in IMO 1979, for carefully reading and editing the early drafts of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

preparation of individuals (or groups of individuals, even whole nations) for future competitions.

In what follows, we give a brief history of contemporary math competitions and present the state of the art in this area. Then we outline how competitions help identify, motivate, and develop higher-ability and talented students. Next we focus on the impact of competitions on education, on educational institutions and on mathematics as a science. Finally, we pose challenges and identify venues for improvement.

## 2. Brief history of mathematics competitions

It is difficult to trace precisely the origins of mathematics competitions for school students; after all, in-class testing (which often resembles small-scale competitions) has accompanied the school system from its very beginning. In fact, the archetype of some competitions can be found outside school, in the society. Newspapers and recreational journals frequently offer prizes for solving crosswords, puzzles, and problems of a deeper mathematical nature. This practice is widely used today by many mathematical journals that publish problems and give awards to school students who provide good solutions.

V. Berinde [2] reports that a primary school math competition with 70 participants was held in Bucharest, Romania, as early as 1885. There were eleven prizes awarded to 2 girls and 9 boys. It cannot be excluded that other competitions were held elsewhere before or after that date too. Nevertheless, the 1894 Eötvös competition in Hungary is widely credited as the forerunner of contemporary mathematics (and physics) competitions for secondary school students. The competitors were given four hours to solve three problems individually (no interaction with other students or teachers was allowed). The problems in the Eötvös competition were specially designed to challenge and check creativity and mathematical thinking, not just acquired technical skills; the students were often asked to prove a statement.

As an illustration, here are the three problems given in the very first Eötvös competition in 1894 (the entire collection of problems and their solutions is maintained by John Scholes at www.kalva.demon.co.uk/eotvos.html):

**P1.** Show that $\{(m, n) : 17 \text{ divides } 2m + 3n\} = \{(m, n) : 17 \text{ divides } 9m + 5n\}$.

**P2.** Given a circle $C$, and two points $A$, $B$ inside it, construct a right-angled triangle $PQR$ with vertices on $C$ and hypotenuse $QR$ such that $A$ lies on the side $PQ$ and $B$ lies on the side $PR$. For which $A$, $B$ is this not possible?

**P3.** A triangle has sides length $a$, $a + d$, $a + 2d$ and area $S$. Find its sides and angles in terms of $d$ and $S$. Give numerical answers for $d = 1$, $S = 6$.

The Eötvös competition model still dominates the competition scene.

The year 1894 is notable also for the birth of the famous mathematics journal *KöMaL* (an acronym of the Hungarian name of the journal, which translates to *High School Mathematics and Physics Journal*). Founded by Dániel Arany, a high school teacher in Györ, Hungary, the journal was essential to the preparation of students and teachers for competitions (about one third of each issue was devoted to problems and problem solving and readers were asked to send solutions). As noted by G. Berzsenyi in the preface of [3], about 120–150 problems were published in *KöMaL* each year; about 2500–3000 solutions were received. The best solutions and the names of their authors were published in following issues. This type of year-round competition helped many young people discover and develop their mathematical abilities; many of them later became world-famous scientists. (For more information, see the journal web site, komal.elte.hu.)

About the same time, similar development occurred in Hungary's neighbor, Romania. The first issue of the monthly *Gazeta Matematică*, an important journal for Romanian mathematics, was published in September 1895. The journal organized a competition for school students, which improved in format over the years and eventually gave birth to The National Mathematical Olympiad in Romania. For legal reasons, the journal was transformed to *Society Gazeta Matematică* in August 1909. The following year, the Romanian Parliament approved the legal status of the new society and this is considered to be the birthday of the Romanian Mathematical Society [2].

What happened in Hungary and Romania in the late 1800's was not something isolated and special to these two countries only; most likely, it reflected a much broader trend. Indeed, international collaboration and solidarity were rising steadily and many national math societies were founded around the same time. The Olympic Games were revived in 1896. The First International Congress of Mathematicians took place in Zürich in 1897. Within several decades, other countries started to organize mathematics competitions. In 1934, a Mathematical Olympiad was organized in Leningrad, USSR (now St. Petersburg, Russia).

## 3. Mathematics competitions today

Today the world of mathematics competitions encompasses millions of students, teachers, research mathematicians, educational authorities, and parents, who organize and take part in hundreds of competitions and competition-like events with national, regional, and international importance every year. Even greater is the number of books, journals, and other printed and electronic resources that help students and their mentors prepare for the various types of competitions.

**3.1. International Mathematical Olympiad (IMO).** Of course, the most important and most prestigious math competition is the *International Mathematical Olympiad* (IMO) – an annual competition for high school students. Directly or indirectly, all other competition activities in mathematics and sciences are related to the IMO.

The idea to organize an international mathematics competition crystallized during the Fourth Congress of Romanian Mathematicians in 1956. Paul Jainta [4] points out that "IMO, the pinnacle of competitions among individuals, was the brainchild of Romania's Tiberiu Roman, an educator of monumental vision." The first IMO took place in Romania (1959) with participants from seven countries: Bulgaria, Czechoslovakia, German Democratic Republic, Hungary, Poland, Romania, and the Soviet Union (USSR). The second IMO (1960) was organized by Romania as well, but since then it is hosted by a different country every year (except 1980, when no IMO was held). Over the years, the participation grew dramatically: the 2005 IMO in Mexico gathered 513 competitors from 93 countries!

Strict formalized rules govern every aspect of the IMO, such as participation, problem selection, assessment of solutions, and distribution of medals (for a description of the IMO, browse erdos.fciencias.unam.mx).

Each country sends a team of up to eight (four in 1982; since 1983, six) high-school students, chaperoned by a team leader and a deputy team leader. The competition itself is held on two consecutive days; each day, the students have four and a half hours to solve three problems. Each year, just before the competition, the six problems are selected by an international jury formed by the national team leaders and representatives of the host country. Even though confined to secondary school mathematics, the problems are rather difficult and solving them requires a significant degree of inventive ingenuity and creativity. Each problem is worth seven points, so the perfect score is 42 points.

Formally, like the Olympic Games, the IMO is a competition for *individuals*; participants are ranked according to their score and (multiple) individual medals are awarded. Nevertheless, again as in the Olympic Games, the medals and points obtained by the participants from each country are totaled and the countries are unofficially ranked, providing grounds for comparison between countries.

The two days of heavy problem-solving are followed by a social program for all the participants. Students get to know each other, discuss alternative solutions to the competition problems, and make plans for their future, while the team leaders share their experiences and best practices in creating new problems and preparing their students for the competition.

With its high standards, the IMO prompts the participating countries to constantly improve their educational systems and their methods for selecting and preparing the students. This yielded a great variety of competitions and mathematical enrichment activities around the world which resists any classification. There are "Inclusive" (open for all) competitions which are intended for students of average abilities, while "exclusive" (by invitation only) events target talented students (a prime example of the second type is the IMO and the national olympiad rounds beyond the first). There are "Multiple-choice" competitions where each problem is supplied with several answers, from which the competitor has to find (or guess, as no justification is required) the correct one. In contrast, "classical style" competitions (like the IMO) require the students to present arguments (proofs) in written form. In "correspondence" com-

petitions, such as those organized by *KöMaL* and *Gazeta Matematică*, the students do not necessarily meet each other, while in "presence" competitions (which form the majority of math competitions) the participants are gathered together, which is believed to provide "equal rights" to all students. There are even mixed-style competitions, with a presence-style first stage and correspondence-style subsequent stages. (We will present some newer styles in more detail later.)

Another indication of the importance of the IMO is the fact that other sciences, such as physics, chemistry, and biology, soon followed suit and started international olympiads of their own. Bulgaria organized the first international olympiads in informatics/computer science (1989) and in mathematical linguistics (2003).

**3.2. Mathematics competition networks.** Like any event with positive social impact, each math competition creates and maintains its network of dedicated people. Numerous math competition journals complement these networks, connecting editorial staff, authors, and readers. Good examples in this direction are Kvant (Russia), Crux Mathematicorum (Canada), Mathematics Magazine and Mathematical Spectrum (UK). The math competition networks range in size from regional to international networks that are associated with large and famous competitions, such as the IMO, Le Kangourou Sans Frontiéres [www.mathkang.org], the Australian Mathematics Competition [www.amt.canberra.edu.au], the International Mathematics Tournament of Towns [www.amt.canberra.edu.au/imtot.html], the Ibero-American Mathematics Olympiad [www.campus-oei.org/oim/], and the Asian-Pacific Mathematics Olympiad [www.cms.math.ca/Competitions/APMO/] – the list is far too short to enumerate all networks that deserve to be mentioned.

The different competition networks are not isolated, as many people naturally belong to more than one network. A different and more formal tie is provided by the *World Federation of National Mathematics Competitions* (WFNMC). The WFNMC was founded in 1984, during the Fifth International Congress of Mathematical Education (ICME5) in Adelaide, Australia. Since then it has a "reserved slot" in the programs of every ICME. Every second year after ICME the WFNMC organizes its own Conference. It has an award, named after Paul Erdös, which is given to people with outstanding contributions to mathematics competitions. The Federation publishes also its journal *Mathematics Competitions* [www.amt.canberra.edu.au/wfnmc.html] which is another powerful tool for networking people engaged with competitions. In 1994 the WFNMC became an Affiliated Study Group of the International Commission on Mathematical Instruction (ICMI), which, in turn, is a commission of the International Mathematical Union [1]. In this way the competitions networks are incorporated into the global mathematical community.

Taken together, these networks form a large global network in the field of mathematics competitions and, more generally, in the classical area known under the (somewhat misleading) name *Elementary Mathematics*. Like in any other area of science, this network operates and lives through its journals, conferences, and workshops, but the periodical regularity of its math competitions adds to its strength and

vitality since the people meet more often. In addition, this global network facilitates the dissemination of best practices in curriculum development and in the work with talented youngsters. New problem solving techniques, new classes of problems, and new ideas about organizing competitions spread quickly around the world. We should not forget also that, through this global network, the Elementary Mathematics (which constitutes an important part of our mathematical heritage) is preserved, kept alive and further developed.

## 4. Why are the competitions needed?

Here is a short and incomplete list of reasons on which we expand later on:

1. higher abilities and talent are identified, motivated and developed;
2. what happens before and after the competition is good for education;
3. talented students are steered to careers in science;
4. competitions raise the reputation of an educational institution.

**4.1. Finding higher abilities and talent.** The educational systems in most countries target mainly students of average mathematical abilities (who form the majority in schools). Additional care is often provided for lower-ability students, so that they could cover the educational standards. The standard curriculum and syllabus requirements pose no significant challenge however to students with higher abilities. They do not feel the need to work hard and, as a result, their mathematical abilities and talent remain undiscovered and undeveloped.

This is a pity, of course, since these higher-ability youngsters are a very important resource for the development of society, provided they are properly educated, motivated, and supported. Unlike other natural resources, such as mineral deposits, which remain preserved for the future generations, if undiscovered and unused, the talent of a young person is lost forever, if it is not identified, cultivated, and employed properly. Competitions and other enrichment activities are obvious remedies for this shortcoming, as they allow students to exhibit their abilities and talent. Moreover, competitions motivate the participants to work hard while preparing for them and, as a result, further develop their abilities and talent.

**4.2. Before and after competitions.** Some opponents to competitions complain that there is no apparent direct connection between the competitions and the mathematics as taught in the classroom. This, in our mind, is a rather narrow approach to the issue. Classroom is only one of the many homes of the educational process. One should take into account the integral impact of competitions and competition-related activities on education. What frequently escapes public attention, which often focuses on a rather small group of happy winners, is the fact that the other, "non-winner" participants,

also gain a lot. While preparing for the competition, and trying to solve the problems during the competition itself, all participants increase their knowledge significantly. Taking into account that in some competitions hundreds of thousands of students are taking part, the integral impact on the learning of mathematics becomes significant for the overall development of the contemporary society. From this point of view the contribution of the International Competition “European Kangaroo" with more than 3 millions of participants is difficult to overestimate.

We should not neglect also what happens in the corridors of the school (or outside the school) after the competition is over. The students are sharing their experiences (successes, failures, new ideas generated, etc.). This has a tremendous educational effect which however is not always given proper attention. The competitions and mathematics enrichment activities can be viewed as events that provide impetus for subsequent discussions among the students (as well as among their friends, parents, etc.). From the viewpoint of acquiring new mathematical knowledge (facts and techniques) these after competition discussions might be as important as the preparation for and the competition itself. Many of us owe a significant part of our knowledge to just such “corridor mathematics". From this point of view the social program after IMO gains additional importance. All this could (and should) have some practical implications for the ways the competitions and other enrichment activities are planned and organized. One should deliberately incorporate possibilities (the more the better) for “after event" discussions, reflections and interactions. There is an unexhausted potential for introduction and sharing new practices in this area.

Finally, while preparing their students for competitions the teachers gain experience how to teach mathematical topics that are currently not in the curriculum. This may become important at later stages, if some of these topics become a part of the school program.

**4.3. Steering talented students to careers in science.** The health and longevity of any social sector depends on how many talented young people are attracted to it. The role of math competitions in identifying talented young people and in attracting them to science should be obvious. Indeed, the fact that a significant number of successful participants in math competitions later become famous scientists was recognized rather early. On 17 July 1929, John von Neumann, who was born in Hungary and was influenced by math competitions, wrote in his letter from Berlin to Professor Lipót Fejér in Budapest ([5]):

> Dear highly honored Professor,
>
> I had the opportunity several times to speak to Leo Szilard about the student competitions of the Eötvös Mathematical and Physical Society, also about the fact that the winners of these competitions, so to say, overlap with the set of mathematicians and physicists who later became well-respected world-figures. Taking the general bad reputations of examinations world-wide into account it is to be considered as a great achievement if the selection works with a 50

> percent probability of hitting the talent. Szilard is very interested in whether this procedure can be applied in the German context and this has been the subject of much discussion between us. However, since we would like above all to learn what the reliable statistical details are, we are approaching you with the following request. We would like:
>
> 1. to have a list of names of the winners and runners-ups of the student competitions,
>
> 2. to see marked on the list those who were adopted on a scientific basis and those adopted for other work,
>
> 3. to know your opinion about the extent to which the prizewinner and the talented are the same people and, for example, what proportion of the former would be worthy of financial support from the State in order to make their studies possible.

Very often the future professional realization of a young person is often predetermined by the "first success." The first area where positive results are achieved often becomes the preferred area in which a person invests time and efforts, which in turn brings more success, stronger motivation, and higher professionalism. Math competitions provide such opportunity for early success and thus help attract good young minds to mathematical and scientific careers. In this way competitions contribute to the development and progress of mathematics and other sciences.

**4.4. Raising the reputation of an educational institution.** The academic reputation of a university depends primarily on the merit of the intellectual achievements of its academic staff. "The higher the reputation of the professors, the higher the reputation of the university" is the essence of this widely accepted belief. What is often overlooked, though, is that the level of the students also has a significant impact on the outcome of the educational process and, in the long run, on the reputation of the institution. While higher-ability students still have the chance of becoming good professionals if trained by ordinary professors, even outstanding professors can fail to produce high-level specialists from mediocre and unmotivated students.

Teachers know well that a few good students in class not only motivate the other students and make them work harder, but also place higher demands on the preparation of the teachers themselves. This two-way challenge influences positively the educational process and improves, directly or indirectly, the reputation of the entire educational institution.

It is no wonder that many universities try hard to attract good students. One of the best ways to achieve this is to organize competitions for secondary school students and to offer incentives, such as stipends or entrance exam waivers, to the winners. Such policies usually yield the expected results, as a special type of relationship develops between organizers and the winners during the preparation for the competition, the competition itself, and the post-competition period, which encourages the winners to

consider seriously (sometimes as the first option) enrolling in the university where the competition (and/or the preparation for it) takes place.

In addition to the obvious advantages, enrolling competition winners has a delayed "value-added" effect to the reputation of a university. After graduation, math competition winners, as people with good problem-solving skills, are more likely to get rapid professional recognition, because they are likely to find solutions to difficult and complex real-life problems easier and faster than others. Once their success is noticed and registered by the working environment, the recognition of the problem-solvers' alma mater increases immediately and almost automatically.

As a success story, consider the University of Waterloo, Canada, and the breathtaking rise of its reputation during the seventies and eighties of the last century. Alongside other plausible explanations, such as good management and excellent academic staff, its success can also be attributed to the fact that the University of Waterloo was the host of the Canadian Mathematics Competition [www.cemc.uwaterloo.ca], which attracted a good portion of the best young minds in Canada.

The William Lowell Putnam Competition, widely known as the "Putnam Exam" and administered by the Mathematical Association of America, is the flagship of annual competitions for university students in North America. While enrolled at the University of Waterloo, the former winners in school competitions performed consistently well in the Putnam Exam, securing a prominent presence of Waterloo in the top five teams in North America. This also was contributing to the reputation of the institution. It is no wonder that, within less than 20 years, the University of Waterloo became one of the leading centers for mathematics and computer sciences in the world.

There is another success story related to the University of Waterloo and the Canadian Mathematics Competition, which shows how a new implementation of an inspiring idea at a new place can yield fantastic results.

The Australian mathematician Peter O'Halloran (1931–1994) spent a part of his 1972–73 sabbatical leave from the Canberra College of Advanced Education (now University of Canberra) at the University of Waterloo. There he gained, as Peter Taylor (Executive Director of the Australian Mathematics Trust) recalls ([6]),

> ... the idea of a broadly based mathematics competition for high school students. On his return he often enthused to his colleagues about the potential value of such a competition in Australia. In 1976, while President of the Canberra Mathematical Association, he established a committee to run a mathematics competition in Canberra. This was so successful that the competition became national by 1978 as the Australian Mathematics Competition, sponsored by the Bank of New South Wales (now Westpac Banking Corporation). It is now well known that this competition has grown to over 500,000 entries annually, and is probably the biggest mass-participation event in the country.

The success of Peter O'Halloran was encouraging for others. André Deledicq started in 1991 the Kangaroo Competition in France (the name reveals the Australian

influence). The Kangaroo Competition is now truly international (albeit with focus on Europe), enjoying more than 3 million participants each year.

It is an appropriate place here to pay tribute to Peter O'Halloran, who had the vision for the future of mathematics competitions and knew the strategies how to achieve the goals. He understood the role of international collaboration in this field and was the major force behind the inception of WFNMC and its association with ICMI as an Affiliated Study Group.

## 5. Competitions and science

Before we go any further, we need to consider a natural question:

> Why are math competitions so good in revealing higher mathematical abilities and inclination to doing research?

The simplest and obvious answer seems to be:

> Because both higher abilities and inclination to doing research are *necessary* to be successful in a math competition.

Necessary, but not sufficient. To be successful in a competition, a student often needs not only a good mind, but a very quick one. Most competitions are limited in time to just 3–4 hours, imposing a significant stress on the nervous system of their participants. Not only do students have to solve the problems correctly, they have to do so quickly and in the presence of their direct competitors. Yet, there are many highly creative students, who do not perform well under pressure. Such "slow thinkers" often come up with new and valuable ideas a mere day (or even just five minutes) after the end of the competition, yet receive no reward or incentive.

Traditional competitions disadvantage such students, even though some of them are highly creative and could become good inventors or scientists. Indeed, what matters in science is rarely the speed of solving difficult problems posed by other people. More often, what matters is the ability to formulate questions and pose problems, to generate, evaluate, and reject conjectures, to come up with new and non-standard ideas. All these activities require ample thinking time, access to information resources in libraries or the Internet, communication with peers and experts working on similar problems, none of which are allowed in traditional competitions.

Obviously, other types of competitions are needed to identify, encourage, and develop such special "slower" minds. The competitions should reflect the true nature of research, containing a research-like phase, along with an opportunity to present results to peers – precisely as it is in real science.

As a matter of fact, such competitions, designed to identify students with an inclination to scientific (not only mathematical) research, already exist. Below we present three of them.

**5.1. Germany/Switzerland.** *Jugend Forscht* (*Youth Quests*) celebrated its 40-th anniversary in 2005. It is a German annual competition for students under the age of 21, who work, alone or in teams, on projects of their own. The projects are presented at special sessions, where the winners are awarded [www.jugend-forscht.de].

Switzerland has a similar competition, which is organized by the *Schweizer Jugend Forscht* (*Swiss Youth Quests*) foundation, established in 1970. The competition, which covers all scientific directions, including social sciences and humanities, has existed since 1967 [www.sjf.ch].

A Google search for the phrase "Jugend Forscht" produced 25 000 hits in 2002; the same search produced half a million hits in 2005! This 20-fold increase speaks for itself, especially since only German language area is included.

**5.2. USA.** Many such programs exist in the USA. As a matter of fact, Jugend Forscht was originally shaped after the many "Science Fairs" in USA. We mention only one such program here, because it emphasizes mathematics and because it was used as a model for similar programs in other countries. The Virginia-based *Center for Excellence in Education* (CEE) was founded by Admiral H. G. Rickover in 1983. It has the following goals [www.cee.org]:

> The Center for Excellence in Education nurtures careers of excellence and leadership in science and technology for academically talented high school and college students. CEE is as well dedicated to encouraging international understanding among future leaders of the world. CEE's programs challenge students and assist them on a long-term basis to become creators, inventors, scientists and leaders of the 21st century.

The major CEE event, sponsored jointly with the Massachusetts Institute of Technology, is the *Research Science Institute* (RSI) [www.cee.org/rsi/]:

> Each summer approximately 75 high school students gather for six of the most stimulating weeks of their young lives. Selected from the United States and other nations, these students participate in a rigorous academic program which emphasizes advanced theory and research in mathematics, the sciences, and engineering.
>
> Students attend college-level classes taught by distinguished professors. Nationally recognized teachers conduct classes designed to sharpen research skills. In addition, students complete hands-on research with top mentors at corporations, universities, and research organizations.

Only outstanding, carefully selected students are admitted to the program. RSI starts with a series of professional lectures in mathematics, biology, physics, and chemistry. The students are paired with experienced scientists and mentors, who introduce them to interesting research topics and share with them the joy and excitement of exploring new territories. The RSI days are filled with research, evening lectures,

ultimate Frisbee, sport events, etc. At the end of the program, the students present their own research, both in written and oral form, and awards are given to the best performers.

The RSI is an international program: almost a third of its students come from other countries. It provides a unique environment for talented students from different parts of the world to meet, live and work together for a relatively long period of time (six weeks seems to be optimal – it is neither too long to become boring nor too short to put unbearable stress). Again, one should not neglect the importance of the networking and friendships fostered by the RSI program for the future development of the participants. The fact that they know each other will make their future collaboration more fruitful. Year after year, the Bulgarian participants in RSI emphasize the social character of the event and the unique atmosphere created during the RSI.

**5.3. Bulgaria.** Before the 1989 political changes, Bulgaria had a venue for talented young people, very similar to the above-mentioned Jugend Forscht and RSI. It was called *Movement for Technical and Scientific Creativity of the Youth* (abbreviated in Bulgarian to TNTM). Students worked on individual scientific projects and presented their work on special sessions, where winners were awarded. Like almost everything else related with the youth, the TNTM movement was under the umbrella of the Young Communists' League (Komsomol). After the democratization of the Bulgarian society, the Communist League disappeared, along with everything related to it, including TNTM.

A decade later it became absolutely clear that actions were needed to revive those activities at the level of contemporary challenges and requirements. The RSI model was adapted to the conditions in Bulgaria and, as one of the "Year of Mathematics" initiatives, the new *High School Students' Institute of Mathematics and Informatics* (*HSSIMI*) was founded in 2000.

Throughout one academic year, the involved high school students (grade 8–12) work on freely chosen topics (projects) in mathematics and/or informatics (computer science). They work individually or in teams and are supervised by a teacher, a university student, a relative, or just any specialist in the field, willing to help. In fact, some recent HSSIMI projects were successfully supervised by former HSSMI participants, who are now university students.

Warmly accepted by the mathematical community in Bulgaria, the HSSIMI organizes three major events: two competition-like sessions and a *Research Summer School*. The sessions are held at the stand-alone *Students Conference* for High School students in January and at the *School Section* of the Annual Spring Conference of the Union of Bulgarian Mathematicians(UBM) in April. The latter section is actually the most visited section at the Spring Conference of UBM, attended by university professors, researchers, teachers, parents, and school peers.

To participate in the HSSIMI sessions, students submit a written paper with the results of their work. Specialists referee the papers, assess the projects, and suggest improvements. Students present their research at the sessions and winners are

awarded. As special award, two of the winners are sent to USA in order to participate in RSI.

The authors of the best projects are invited to a three-week *Research Summer School*. During the first two weeks, eminent specialists from universities, research institutes, and software companies give lectures and practical courses in mathematics and informatics. As in similar programs, the main goal of this preliminary training is to expand the students' knowledge in topics of their interest and to offer new problems for possible projects. During the third week, students hold a High School Students Workshop, where they briefly present their ideas for new projects.

For the short period of its existence, the HSSIMI became a valuable addition to the established (and rather densely populated) system of traditional competitions in Bulgaria. As was planned and expected, the HSSIMI attracted students who were not regulars in the traditional competitions.

Similar initiatives can be found in other countries. There are positive signs of networking between them as well. Good examples in this direction are the Tournament of Towns Summer Conference and the annual International Mathematics Projects Competition (IMPC) in Kazakhstan. Reflecting more closely the nature and spirit of research process, these kinds of activities also attract excellent minds to mathematics and definitely deserve better recognition and support by the professional mathematical communities around the world.

## 6. What to do next?

In addition to enhancing the traditional math competitions and developing the non-traditional initiatives discussed above, there are other venues for future improvements, such as implementing the current science trends into competitions, targeting other audiences, and supporting and developing the human resources standing behind competitions and other related activities.

**6.1. Algorithms in mathematics.** The nature of mathematical research has changed significantly since considerable computing power came to the desk of almost every researcher and student. Mathematicians today can conduct complicated numerical experiments, use software for complex algebraic and analytic transformations, find patterns in huge data sets. Like the experiments in other sciences, this could help reject some conjecture or formulate a new one. Thus, research in mathematics became similar to research in the other sciences.

All this is based on mathematical algorithms. Algorithmic thinking is getting higher importance and successfully complements the "axiomatic" approach and thinking in mathematics.

This change should be duly reflected in the creation and selection of competition problems. Perhaps more problems should be offered at various competitions where algorithms and their properties are focused in order to cultivate algorithmic thinking.

Otherwise, we will become witnesses of a "brain-drain" and the best young minds will be driven to competitions in informatics.

**6.2. Teamwork.** Working in teams is a well-established trend in modern science. For centuries, research in mathematics has been a solitary endeavor. Today, we see more and more teamwork in mathematics and, especially, in its applications. This reveals yet another similarity between modern mathematics and the other sciences (where teamwork has traditionally deeper roots). The ability to work in a team is valuable skill that could and should be cultivated early on.

Mathematics team competitions could contribute a lot in this direction. There are many such competitions around the world; it only makes sense to make them more popular.

**6.3. Competitions for university students.** Even though they are not the focus of this paper, mathematics competitions for university students, among other virtues, help attract talented young people to academic careers in mathematics.

Some of these university-level competitions are highly respected and have existed for many years. The above-mentioned Putnam Exam is more than 65 years old. The International Mathematics Competition for University Students began in 1994 [www.imc-math.org]. Of course, there are many other such competitions, but their number is still much smaller than the number of competitions for secondary school students, providing plenty of opportunities for new initiatives and international collaboration in this area.

**6.4. Teachers and the competitions.** In many countries, year after year, some schools consistently "produce" more competition winners than other schools. What is the reason behind this phenomenon? Why are some schools more successful than others?

The reasons may be numerous and fairly different in nature. Very often, however, the prominent success of a particular school can be attributed to the dedicated efforts of a single teacher or a small group of teachers. For these excellent teachers, teaching is a vocation, a mission, and not just means to make both ends meet. Such special teachers are real assets for the school and for the whole country. They possess both the necessary scientific ability and the extraordinary personality needed to identify and motivate for hard work the future winners in competitions.

Such teachers need special care, though. Their higher scientific ability is acquired very slowly, at the expense of great personal efforts. It is no secret that the success of these teachers depends very strongly on their working environment and on the appreciation by their colleagues and administration. Very often however the actual working conditions in the schools do not support the work and the development of these dedicated teachers.

There is a lot that can (and have to) be done in order to improve the situation. For instance, the materials available to the teachers should not include problems and

solutions only, but also provide didactic instructions for the teachers how to use these materials in their work with higher ability students and what type of reactions and difficulties to expect on the side of students. For this to happen a special research is needed, conducted with the help of professional math educators.

Many organizations which are involved with competitions are also organizing seminars and workshops for teachers. There is a valuable experience in many countries in the work with such teachers. The positive results and the problems could be discussed and evaluated with the aim of disseminating the good practices. Teachers are the major human resource for the development of competitions and related activities.

Another problem is that often competition-like activities are not "at home" (and therefore not appreciated) both in Mathematics Departments (because "they concern Elementary Mathematics") and in Mathematics Education Departments (because they are "too mathematical and refer to the relatively small group of talented students"). It is time for both communities (research mathematicians and mathematics educators) to understand their joint interest in supporting competitions and competitions – related activities.

## 7. Summary

Competitions have influenced positively mathematics education and its institutions in different ways for more than a century.

Engaging millions of students and educators, math competitions have a distinguished way to identify, motivate, and develop young talent, steering it to careers in science.

Mathematics competitions have matured and formed an immense and vibrant global network which contributes significantly to the preservation and the maintenance of mathematical heritage.

The flagship IMO not only serves as the "golden standard" for numerous other competitions in mathematics and the sciences (especially with its often-overlooked social program), but it also provides a constant stimulus for improvement of school systems around the world.

Traditional competitions are complemented by more inclusive and less known events that emulate more closely real research and engage even broader student audience.

Nevertheless, stronger consolidation and collaboration of teachers, schools, universities, and educational authorities is needed in order to meet the challenges of the new century.

## References

[1] Bass, H., Hodgson, B. R., The International Commission on Mathematical Instruction. *Notices Amer. Math. Soc.* **51** (2004), 639–644.

[2] Berinde, V., *Romania – The Native Country of International Mathematical Olympiads*. A brief history of Romanian Mathematical Society. CUB PRESS 22, Baia Mare 2004.

[3] *Century* 2 *of KöMaL*. V. Oláh (editor), G. Berzsenyi, E. Fried and K. Fried (assoc. editors), KoMaL, Janos Bolyai Mathematical Society/Roland Eotvos Physical Society, Budapest 1999.

[4] Jainta, P., Problems Corner: Contests from Romania, *EMS Newsletterr* **35** (2000), 20–24; www.emis.de/newsletter/newsletter35.pdf.

[5] von Neumann, J., Letter to Professor Lipót Fejér in Budapest (1929). In *Neumann Archive*, OMIKK, Budapest; www.kfki.hu/fszemle/archivum/fsz9905/papp.html.

[6] Taylor, P., Obituary: Peter Joseph O'Halloran (1931-1994), *Mathematics Competitions, Journal of WFNMC* **7** (2) (1978), 12–17; www.amt.canberra.edu.au/obitpoh.html.

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Akad. G. Bonchev Street, Block 8, 1113 Sofia, Bulgaria
E-mail: kenderovp@cc.bas.bg

# Understanding and misunderstanding the Third International Mathematics and Science Study: what is at stake and why K-12 education studies matter

Alan Siegel*

**Abstract.** The technical portion of this paper concerns a videotape classroom study of eighth grade mathematics lessons in Japan, and how methodological design errors led to conclusions that are refuted by the actual video data. We document these errors, and trace their distillation into one- and two-sentence education policy recommendations articulated in U.S. government position papers, implemented in classrooms across the U.S. and imported by countries around the world. We also present the historical context needed to understand the misrepresentations cited in support of questionable education policy.

**Mathematics Subject Classification (2000).** Primary 97D40; Secondary 97-02.

**Keywords.** Third International Mathematics and Science Study, TIMSS.

## 1. Introduction

The outstanding results for the top-performing countries in the Third International Mathematics and Science Study (TIMSS) have generated widespread interest in best teaching practices around the world. In the TIMSS Videotape Classroom Study by James Stigler et al. [31], the teaching styles in Germany, Japan, and the U.S. were compared in an effort to discover what makes some programs so successful. The conclusions from this comparison are striking and have been widely cited, but often in a highly trivialized and even inaccurate manner. Moreover, this particular study, as we will show, is marred by design errors that raise serious doubt about some of its most influential conclusions. Indeed, it is these very findings that have been cited and accidentally distorted in support of the latest reform programs and education policies – both in the U.S. and elsewhere.

For example, it is widely acknowledged (cf. [31, p. 134]) that Japanese lessons often use very challenging problems as motivational focal points for the content being taught. According to the Glenn Commission[1] Report [10, p. 16],

*The author is grateful to the ICM, NSF and AMS for their support of this presentation. Disclaimer: although the assessments and statements in this paper have been made in good faith by the author, they should not necessarily be viewed as representative of or endorsed by the ICM, NSF or AMS.

[1]The commission's proper name is the National Commission on Mathematics and Science Teaching for the 21st Century. It was chaired by former U.S. Senator and astronaut John Glenn. The year-long Commission was

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

> "In Japan, . . . closely supervised, collaborative work among students is the norm. Teachers begin by presenting students with a mathematics problem employing principles they have not yet learned. They then work alone or in small groups to devise a solution. After a few minutes, students are called on to present their answers; the whole class works through the problems and solutions, uncovering the related mathematical concepts and reasoning."

We revisit the TIMSS Videotape Classroom Study to resolve the one crucial classroom question that both the Glenn Commission and the TIMSS Classroom Study group failed to address:

> How can Japanese eighth graders, with just a few minutes of thought, solve difficult problems employing principles they have not yet learned?

We will see that the technique required to solve the challenge problem of the day will have already been taught, and that the lesson begins with a review of the fundamental method needed to solve the problem. Students begin working on these problems individually – not in groups. Sometimes group-work is allowed for second efforts on a given assignment, but only after individual seat-work. These lessons include student-presented solutions, but the presentations are closely supervised by the teacher, and the time allocated for this activity is limited so that students will be able to work on a second challenge exercise of the same type, and the teacher will have enough time to show how to apply a fundamental technique as many as ten times – all in a single lesson. Stigler's videotapes reveal master teaching of substantial content hidden within a warm and inviting teaching style. Students do indeed participate, but in moderation, and subject to the vigilant oversight of instructors who ensure that no one wanders off course.

It is also worth noting that the Videotape Classroom Study identified some of the significant differences between the current reform positions and Japanese teaching practices. For example, it pointed out that students did not use calculators in the Japanese classes, and that Japanese teaching has a far higher concentration of proofs and derivations than both reform and traditional programs in the U.S. The Videotape Study also found that Japanese teachers spend more time lecturing than even traditional U.S. teachers.

These distinctions not withstanding, the notion that Japanese teaching might be implementing U.S. reforms is given far greater emphasis in a major Government report, which flatly declares:

> "Japanese teachers widely practice what the U.S. mathematics reform recommends, while U.S. teachers do so infrequently [25, p. 9]."

---

mandated to develop a strategy to raise the quality of mathematics and science teaching in all of the nation's classrooms. Unfortunately, the cited quote was, quite possibly, the most substantive paragraph in their report to the nation. The preliminary Glenn Commission report cited Stigler and his TIMSS Videotape Classroom Study as the source of this finding, although the final version omitted the specific citation.

This report on best teaching practices worldwide makes no mention of any differences between the U.S. reforms and Japanese teaching styles. Evidently, its perspective (see also [25, pp. 40–43]) differs from that of its source of primary information, which is the more cautiously worded TIMSS Videotape Study [31]. Moreover, the differences identified in the Videotape Study – which concern direct instruction, calculators, and teacher-managed demonstrations – are all matters of contention in the U.S. debate over classroom reform.

Finally, we note that studies of individual classroom lessons – no matter how comprehensive – are necessarily incomplete. They cannot detect how coherent a curriculum might be day-by-day, much less over the course of years, and are ill-equipped to assess the completeness of a given math curriculum.

## 2. Background

The need for sound – and indeed first-rate – K-12 mathematics programs is well understood. In the U.S., many reform programs have been implemented over the last fifty years, but the evidence shows that on balance, we have made very modest progress toward this goal of world-class math education.

The majority of our past reform efforts can be characterized as a tug of war between traditional and student-centric education movements. Just one of these programs was sufficiently different to deserve special mention: the so-called New Math that originated in the 1950s, and which was widely implemented in the '60s. This reform was pioneered by mathematicians, and was the only program ever to attempt to teach elementary mathematics from an informal set-theoretic perspective. It failed, in part, because its implementations did not provide safeguards to ensure that mainstream American students – and teachers – could handle the material, which is an error that the current reformers have been very careful to avoid. Finally, the program has historical importance because its failure led to a fairly sharp separation between those concerned with K-12 math education and those interested in mathematics research and college teaching.

In the mid-1980s, a new version of student-centric learning and teaching began taking hold in the mathematics education community, and it is fair to say that these ideas have swept the American schools of education, and are likewise well represented by advocates in many other parts of the world.

In 1989, these ideas were codified into teaching policy when "educators ... carefully articulated a new vision of mathematics learning and curriculum in the National Council of Teachers of Mathematics' (NCTM's) *Curriculum and Evaluation Standards for School Mathematics* [6]." The 1989 Curriculum Standards [20], together with the follow-up 1991 Teaching Standards [21] and the 1995 Assessment Standards [22] called for a redirection of focus from what to teach grade by grade to new ideas about how to teach and how to assess student progress. And with the publication of these documents, the NCTM completed its transformation from an organization that

began in the 1920s with ties to the Mathematical Association of America, and that had been led by content-oriented math teachers who endorsed the revolutionary New Math of the '60s, to an organization led by professors of mathematics education who endorsed a new type of revolutionary math program[2] in the '90s.

Loosely put, the theoretical core of this new vision of education is called constructivism. Like most complex social theories, constructivism is founded on a few main principles, has many interpretations and derived consequences, and a bewildering variety of implementations. A thumbnail (and necessarily incomplete) sketch of the main principles of constructivism is as follows.

The philosophical basis of constructivism is that everyone learns differently, and that we learn best by integrating new knowledge into our own core understandings and thought processes. Therefore, education is most effective when it engages the learner to become the main agent in the learning process. That is, learning should be engaging in every sense of the word. Since we learn by discovering and by doing, learning is a quintessentially social process wherein through mutual interaction, we organize, communicate, share, and thereby develop deepened understanding. Moreover, content should be based on real-world problems to reach each learner's core knowledge base, and to maximize the purposefulness of each lesson.

As stated, these objectives have merit – especially for teaching younger learners. Indeed, the author believes that the debate over abstract constructivism misses the point. However, the teaching reforms advocated by the NCTM include, in addition to abstract principles, very applied recommendations that have significant impact on curricula, pedagogy and the opportunity for students to learn mathematics.

Thus, the real questions concern the content and training provided by the reform program implementations, as well as the consequences of the derivative theories of learning and testing that are put forward as logical consequences of constructivist principles. And it is this debate about what kinds of education programs work that defines the context for the TIMSS Videotape Classroom Study and the classification of Japanese pedagogy.

**The impact of reform principles on classroom structure and course content.** The applied education theories advanced by contemporary reformers must be sketched out if the various assertions about Japanese teaching and the latest reform recommendations are to make sense.

The principle of discovery-based learning aims to have the students themselves discover mathematical principles and techniques. According to Cobb et al. [5, p. 28],

[2] The NCTM reform program was also endorsed by the federal department of Education and Human Resources, which provided funds to create reform-compliant textbooks, to support their use, and to support studies designed to prove that the new programs were effective. To date, more than $75 million has been allocated to produce these new mathematics textbooks, and about $1 billion has been spent on programs to foster their use. The Educational Systemic Reform programs, for example, ran for nine years with an annual budget of about $100 million, and related programs for K-12 math and science education received comparable funding. More about the history of these programs can be found in [32].

> "It is possible for students to construct for themselves the mathematical practices that, historically, took several thousand years to evolve."

In the 1999 Yearbook of the National Council of Teachers of Mathematics, the article "Teaching Fractions: Fostering Children's Own Reasoning" by Kamii and Warrington [15] advises:

> "1. Do not tell children how to compute by using numerical algorithms. . . .
>
> 2. Do not tell children that an answer is right or wrong. . . .
>
> 3. Encourage children to use their own reasoning instead of providing them with ready-made representations or 'embodiments.'
>
> 4. Ask children to estimate solutions to problems first because estimation is an effective way to build strong number sense."

To be fair to the authors, it should be pointed out that they provide alternatives to prohibitions 1, and 2. For example, they recommend that the issue of correctness be resolved by the entire class through cooperative discussion.

These discovery-based policies are often implemented via the *workshop model* of teaching where students are seated in clusters of four desks facing each other with no central lecture place in the classroom. This organization is designed to foster collaborative learning and to reinforce the teacher's role as a "guide on the side" as opposed to the "sage on the stage." In some programs, the purpose of the teacher is to introduce the exercise of the day. The students then work in groups of four to discover what they can about the problem. In the next phase, the students present their findings to the class, and an active discussion typically ensues. The teacher might have a role that is confined to being a moderator to maintain order in the discussions. Likewise, some of the programs feature unsupervised group-work with the teacher serving mainly as a passive observer.

In the higher grades, the U.S. discovery-based programs feature markedly diminished content depth, and the project-based texts exhibit poor coherence in their management of topics and offering of reinforcement exercises. To date, some reform programs simply omit material that does not fit within this model. Moreover, this style of teaching, absent sufficient guidance from the teacher, is typically very time consuming, and the slow pace cannot help but limit the curriculum.

For example, on page 315 of a tenth grade reform geometry textbook [37], exercise 24 asks the student to draw an equilateral, an isosceles, and a scalene triangle, and to draw the medians and observe the outcome in each case. The assignment also asks the students to measure the lengths of the medians and the distance from the vertices of each triangle to its centroid. The problem finishes by asking, "What do you conclude?" No proofs are offered or requested, and for good reason. The study of similar triangles begins in Chapter 13 on page 737, where the final two chapters of the book present content that is less observation-based.

In 2001, I was invited to observe some of these workshop model classes at a magnet high school in lower Manhattan. In one of the ninth grade classes, the lesson

problem of the day was (in mathematical terms) to determine the equation of a line through the origin that does not intersect any additional points on the integer Cartesian lattice in $\mathbb{R}^2$. The students began the exercise working unsupervised in groups of four. Then the class convened as a whole to discuss their findings with the teacher serving as moderator. The tenth (or so) student to speak observed that if the line were to intersect another lattice point, then it would have a rational slope. The teacher then called on another student, and this key observation was soon lost. The discussion devolved into an unsuccessful effort to understand the difference between rational numbers with finite decimal representations and those with repeating decimal expansions, and the math period ended with no solution to either question.

In a televised eleventh grade lesson [24] from a reform textbook series [7], students seated in groups of four were given the following problem. The teacher displayed boards of different lengths, widths, and thicknesses suspended between pairs of bricks. A karate expert, he explained, can deliver the tremendous energy necessary to break a strong board. For the first part of the lesson, the students were asked to determine a formula for the energy necessary to break a board as a function of its length and thickness. The students discussed the question with great enthusiasm. There was no evidence of any physical modeling, and it was not clear if the class knew Hooke's law or not. In the second portion of the lesson, a representative from each group presented the group's thoughts to the class. The first to speak was able to intuit that a longer thinner board would be easier to break, but nevertheless went on to opine that the formula for the energy $E$, as a function of the length $L$ and thickness $T$, should be $E = L + T$. Another group thought that the formula should be $E = kLT$, where $k$ is a constant that depends on the physical properties of the wood. In the next portion of the lesson, students were given strands of dried spaghetti to form a bridge between two tables, pennies to use as weights, and a paper cup plus paper clip to suspend on the strand(s) of spaghetti. They then conducted tests with different lengths and strand counts to see how many pennies were necessary to break the spaghetti – thus measuring the breaking force, which was misrepresented as energy. The use of multiple strands served to emulate different thicknesses (albeit incorrectly). Data was gathered for 1 to 5 strands, and distances of 2 to 5 inches. Then the students used their graphing calculators under the supervision of the teacher to determine the best fit for the data, which was $E = 10\frac{T}{L}$, where $E$ is measured in pennies, $T$ in spaghetti strands, and $L$ in inches.

The TV program closed by noting that with the introduction of this new curriculum, grades were higher, and more students were electing to take three and four years of math classes. Of course, the stacking of spaghetti strands to model thicker pasta constitutes a fundamental conceptual error. It is no accident that plywood is manufactured with bonded layers, and as straightforward mathematical modeling shows, strength, in a simple model of deformation, is proportional to the square of a beam's thickness. Likewise, the confusion between force and energy ill serves the students, as does the lesson's implication that mathematics might be an experimental science.

The reforms seeking to maximize engagement include mandates to avoid drill and – by extension – the kind of practice necessary to instill knowledge transfer to long-term memory.

In concrete terms, the reform programs do not teach the multiplication table in elementary school. Ocken reviewed all of the printed materials produced by one of the elementary school reform programs [33] for grades K-5, and found fewer than 30 problems asking students to multiply two whole numbers, both of which contain a digit greater than five[23]. This program implements the reduced emphasis on pencil and paper calculations as recommended in the 1989 NCTM Standards, and, as recommended, supports student work with calculators even in the earliest grades.

Likewise, the standard place-based rules for multiplication of multidigit integers are no longer taught as essential material. Opponents of these reforms see the structure of place-based multiplication as precursor knowledge that helps the learner internalize the more abstract operations of polynomial arithmetic.

In one textbook series [16], the division of fractions was simply omitted from the curriculum. And long division is long gone from these programs.

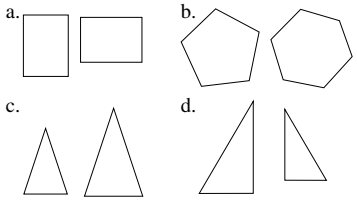
To maximize engagement, reformers recommend that problems and content be *situated*, which means that exercises, derivations and even theorems should be presented in an applied context whenever possible. More generally, abstraction and symbolic methods are eschewed. Of course, the foregoing comments about abstraction and symbolic methods are just words. In order to understand them, we again take a few quick peeks into the reform textbooks to see how these theories and recommendations are turned into practice. For example, one ninth grade reform book [8] has, scattered among its 515 pages, only 25 pages that even contain an equal sign. Of these, only pages 435 and 436 actually concern algebra. The totality of the information about algebra is on page 436, and is as follows.

> "Some such equations are easier to solve than others. Sometimes the particular numbers involved suggest tricks or shortcuts that make them easy to solve. In each of the equations below, the letter x stands for an unknown number. Use any method you like to find the number x stands for, but write down exactly how you do it. Be sure to check your answers and write down in detail how you find them.
>
> $$\frac{x}{5}=7 \qquad \frac{x}{6}=\frac{72}{24} \qquad \frac{x}{8}=\frac{11}{4} \qquad \frac{x}{7}=\frac{5}{3}$$
> $$\frac{x+1}{3}=\frac{4}{6} \qquad \frac{5}{13}=\frac{19}{x} \qquad \frac{2}{x}=6 \qquad \frac{9}{x}=\frac{x}{16}$$"

The preference for encouraging ad hoc "tricks" and "shortcuts" instead of teaching systematic methods is evident. Indeed, the text does not present any methods for solving these problems. The passage also illustrates how the these new programs encourage students to write expository explanations and avoid teaching students to

develop and record logical solution strategies based on correct operations, problem decomposition, and the layered application of systematic methods.

On page 416 of this ninth grade text, problem 3 reads as follows.




3. Consider the following pairs of figures. In each case, state whether you consider the shape to be the same or not, and why.

The chapter goes on to explore some of the most elementary properties of similarity, but the development is probably closer to the level of sixth grade than ninth, and the overall content of the textbook is far weaker than, say, the standard sixth grade books used in Singapore [13].

The comparison with the Singapore books is worthy of elaboration. In an American Educator article [1], the mathematician Ron Aharoni writes about what he learned using the Singapore math program to teach first grade in Israel. He points out that these lessons encourage students to describe problems in words, and feature more discussion than is common in traditional programs. These characteristics are consistent with some of the constructivist principles. There are, however, fundamental differences between this teaching style and the applied recommendations and prohibitions that characterize – and indeed define – the latest reform practices. Aharoni describes how he actively teaches insights based on his mathematical knowledge – even in first grade. And he also points out that significant reinforcement is necessary to help first graders integrate this first grade content into their own thinking. Interestingly, the fifth and sixth grade Singapore texts [13] exhibit a transition from this verbal/expository approach of reasoned problem representation to an informal but precise prealgebra. The books present – with many detailed examples – a kind of pictorial algebra, where a physical segment might be used to represent an unknown length. The modeling allows graphical unknowns to be added, subtracted, and multiplied and divided by integers in physical representations of equations. Students solve many carefully constructed word problems with this modeling process and its physical representation of variables. This representation is used to strengthen intuition and understanding as preparation for variables and algebra. By the sixth grade, the students are using the method to solve sophisticated word problems that would challenge U.S. high schoolers. No U.S. reform text presents such a coherent curriculum, and none provides a systematic increase of content and problem depth chapter-by-chapter and over the course of years to build deepening layers of understanding on behalf of the learner.

In terms of pedagogy, Aharoni emphasizes the importance of deep content knowledge and a deep understanding of *what* is being taught as prerequisites for deciding *how* to teach a particular topic [1, p. 13]. He says that the understanding of fundamental mathematical principles can be taught, but this instruction requires active teaching by a very knowledgeable teacher.

The current reform programs, by way of contrast, aim to teach less, not more. In a ninth grade reform algebra text, for example, students receive enough training to solve for $x$ in the equation $y = 3x + 2$, but there is just one equation in the book that uses variable coefficients. This one exception, which is on page 748 reads [9]:

> "Show how to derive the quadratic formula by applying completing the square to the general quadratic equation, $ax^2 + bx + c = 0$."

This question requires a tremendous leap in skill given the text's limited use of equations with variable coefficients. Moreover, the presentation on completing the square is so weak that it is inconceivable how any but the most exceptional student could learn enough to solve this problem. The totality of the exposition reads:

> "Here's an example of how to use completing the square to solve the quadratic equation $x^2 + 6x - 2 = 5$.
>
> Since $-2$ doesn't make $x^2 + 6x$ a perfect square, it is in the way. Move it to the other side: $x^2 + 6x = 7$.
>
> Add 9 to both sides to make the left side a perfect square: $x^2 + 6x + 9 = 16$.
>
> Write the left side as a perfect square: $(x + 3)^2 = 16$."

There is no attempt to teach a systematic approach for completing the square, or to explain how the magical 9 was selected for use in this particular case.

The avoidance of abstraction and symbolic coefficients, and the recommendations against teaching systematic methods have undermined the quality of the textbook. This instance of teaching by one explicit example cannot instill wide-spread understanding. And the inclusion of the exercise to derive the quadratic formula (which is just about the last problem in a very long book) would appear to be based less on it being an appropriate exercise than on the need to include the topic in the curriculum.[3]

Ralston recommends the outright abandonment of pencil and paper calculations in favor of mental arithmetic supplemented by calculators [26]. Non-reformers disagree, and suggest that proficiency in arithmetic is not taught for its own sake but rather to strengthen the learner's core knowledge and intuition as a prerequisite for understanding fractions. Arithmetic fluency is even more important for a mastery of and fluency in algebra, where the rules of arithmetic are revisited at an abstract level with the introduction of variables and exponents. Many teachers report that those who lack a grounding in the concrete operations of arithmetic experience great difficulty with algebra and its manipulation of symbols. Other non-reformers argue that the written record of pencil and paper problem solving documents a student's approach to a problem, which can be reviewed by the student and the teacher for conceptual errors as well as computational mistakes. Non-reformers also argue that it is the use of the written record that allows learners to combine fundamental steps into more

[3] It is also fair to say that some of the most project-based reform texts are designed around sequences of typically unrelated projects, which result in a disorganized and incomplete curriculum with very few review and reinforcement exercises (cf. [33], [8], [7], [9]).

complex solutions that are too detailed to retain as mental calculations. In addition, it is argued that the written representations of algebra bring a precision of expression, of computation and of modeling that surpasses the written word in accuracy, clarity, and simplicity.

The purpose of this inside review of American mathematics education was to identify the controversies arising from the latest reforms in concrete (i.e. situated) – as opposed to abstract – terms. It is time to explain why Japanese pedagogy has become a topic of worldwide interest, and to investigate how well it aligns with the latest reform principles.

**The Third International Mathematics and Science Study.** TIMSS is an enormous umbrella project that seeks to measure academic achievement around the world, and which includes many subsidiary studies that analyze a host of related issues in an effort to determine how best to improve math and science education. TIMSS began in 1994–95 with the testing of 400,000 students worldwide at grades four, eight, and twelve. It has grown into a quadrennial program that conducted additional testings and data acquisitions in 1999 and 2003, and has already begun to lay the groundwork for the next round in 2007. The program now includes nearly fifty countries, and the studies cover a large number of independent projects with publications in the many thousands of pages.

Although there have been some fluctuations in the TIMSS rankings over the last decade, and the participating countries have varied to some degree over time, the overall results have been much more consistent than not. This fact is probably a testament to the meticulous effort to maintain balanced student samples from the participating countries, and the care that is exercised in the testing protocols and data analyses. The project also deserves very high marks for adhering to a wonderfully high standard of scholarship. The research projects produce not only reports of findings but also detailed documentation of the data acquisition and analysis procedures and indeed every aspect of project methodology. When feasible, these studies even publish enough raw data for independent researchers to review every step of the research effort for independent assessment.

Despite the wealth of information provided by the TIMSS publications, it is fair to say that two specific TIMSS findings have captured the majority of the headlines, and have had the greatest influence on classroom practice and education policy.

The most eye-opening results come from the achievement scores of students around the world. For example, in the little multicultural, multilingual, top-performing country of Singapore, some 46% of the eighth graders scored in the top 10% of the world. And 75% of their students placed among the top 25% of all eighth graders worldwide. Just 1% of their students placed among the bottom 25% of all eighth graders around the world. This is a stunning achievement. Singapore has indeed shown what it really means to have an education system where no child is left behind.

Moreover, these performance results have held up with remarkable consistency in each of the TIMSS testing rounds. Just a notch down from Singapore, the next group

of top performers have been Korea, Hong Kong, Chinese Taipei (formerly known as Taiwan) and Japan (mostly in this order) with Flemish Belgium trailing somewhat behind, but consistently next in line.

The U.S. scores are also worth mentioning. Roughly put, American fourth graders and eighth graders scored somewhat above the international average. But at the twelfth grade, the U.S. scored at the bottom of the industrialized world, and only significantly out-performed two countries: South Africa and Cyprus. No other country fell so far so fast. There was also a more sophisticated twelfth grade test that was reserved for twelfth graders in advanced math programs in the participating countries. On that test, the U.S. was next-to-last; even Cyprus performed significantly better.

For completeness, it should be noted that the twelfth grade testing has not been repeated since 1995 and the U.S. plummet just described. This is unfortunate because the lack of follow-up testing forces us to infer whether the American mathematics programs have recovered from the results documented in 1995. Moreover, the real purpose of a K-8 program is to prepare students for subsequent study as opposed to an eighth grade TIMSS test. So our understanding of mathematics education around the world would be greatly enhanced by a schedule of testing that includes grade twelve as well as grades four and eight.[4]

In view of the absence of follow-up twelfth grade testing, one could speculate that the American TIMSS scores might show that the newest programs are beginning to make a difference. After all, the latest math reforms are often introduced at the earlier grades first, and then extended by one grade level per year. Could it be that U.S. high school students are performing better now because more of them are participating in reform math programs? The answer seems to be a clear no. A variety of studies[5] have documented very little progress in high school math achievement over the last decade. To date, the NAEP scores, for example, have been most notable for their lack of improvement.

In short, TIMSS testing shows that the US, and indeed most of the world have K-12 mathematics programs that are nowhere near the quality of the best programs worldwide. These results constitute a compelling argument for continued testing on an international scale. Simply stated, TIMSS is one of our best mechanisms for identifying unforeseen weaknesses in national programs, and for discovering exemplary programs that can be investigated in an effort to improve domestic teaching.

The other finding that has generated enormous impact can be traced to "TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States" [31]. For convenience, we condense the TIMSS Videotape Classroom Study's name to TVCS.

[4]For countries such as Singapore, which do not have a twelfth grade, the testing might well be given at the completion of the secondary education system.

[5]See, for example, Too Little Too Late: American High Schools in an International Context by William H. Schmidt. In *Brookings Education Policy papers 2003* (ed. by Diane Ravitch), pp. 253–277.

**The Videotape Classroom Study documentation.** During 1994–95, the TVCS team recorded 231 eighth-grade mathematics lessons in Germany, Japan and the U.S. The TVCS project report by Stigler et al. [31] contains an extensive analysis of these tapes and a description of the data acquisition and analysis methodologies. Stigler and James Hiebert subsequently conducted a joint study of Japanese training in pedagogy, which has strong cultural traditions that are surprisingly different from the programs of teacher development in the U.S. [30]. In 1999, Hiebert and Stigler began a second TIMSS videotape classroom study [11] that covered a broader selection of higher performing countries.

These videotape study projects produced a variety of supporting documentation [34], [35], [36], [14], [12], but the follow-up study did not record a new series of Japanese lessons and instead relied on the earlier tapings. We cover the main findings from the second study and the differences in its methodology and conclusions (which may well have resulted from criticisms of the earlier project), but will focus primarily on the 1995 TVCS, which remains the far more influential of the two publications.

The 1995 project produced a publicly available videotape [34] that begins with Stigler presenting an overview of the Japanese lessons that is very similar to the description already quoted from the Glenn Commission Report. It then shows carefully selected representative excerpts of the geometry and algebra lessons recorded in Germany, Japan, and the U.S. The German and American lesson samples were produced in addition to the original 231 recordings, which are not in the public domain due to confidentiality agreements. The Japanese excerpts were selected from the original 50 tapings recorded in Japan, and disclosure permissions were obtained after the fact.

The TIMSS videotape kit includes a guide to the excerpts [36] and a CD ROM [35] is available with the same excerpts, but without Stigler's introduction.

## 3. What the Japanese video excerpts show

**Geometry.** The tape shows the Japanese geometry lesson beginning with the teacher asking what was studied the previous day. After working to extract a somewhat meaningful answer from the class, he himself gives a summary: Any two triangles with a common base (such as $AB$ in Figure 1) and with opposing vertices on a line parallel to the base (such as the line through $D$, $C$ and $P$) have the same area because the lengths of their bases are equal, and their altitudes are equal. The teacher states this principle and uses his computer graphics system to demonstrate its potential application by moving vertex $P$ along the line $CD$. The demonstration

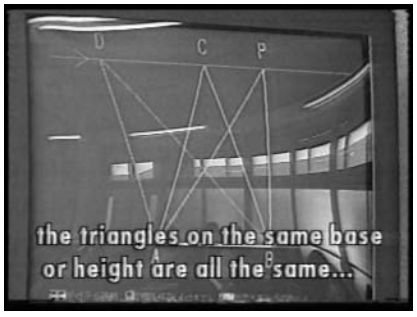



Figure 1

shows how to deform triangle $ABP$ in a way that preserves its area. Next, he explains that this principle or method is to be the "*foundation* [36, p. 136]" for the forthcoming problem, which he then presents. It is the following.

Eda and Azusa each own a piece of land that lies between the same pair of lines. Their common boundary is formed by a bent line segment as shown.

The problem is to change the bent line into a straight line segment that still divides the region into two pieces, each with the same area as before.

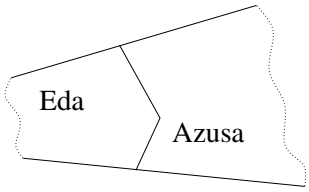



Figure 2

Despite the previous review, the problem is still going to be a challenge for eighth graders, and it is fair to infer that the teacher understands this very well. In geometry, one of the most difficult challenges in a construction or proof is determining where to put the auxiliary lines. These lines are needed to construct the angles, parallel lines, triangle(s), etc. that must be present before a geometry theorem or principle can be applied to solve the problem. For the exercise in Figure 2, the key step is to draw two crucial auxiliary lines. One defines the base of a triangle that must be transformed in a way that preserves its area. The other is parallel to this base, and runs through its opposing vertex.

So what should a master instructor do? The answer is on the tape.

After explaining the problem, the teacher asks the students to estimate where the solution line should go, and playfully places his pointer in various positions that begin in obviously incorrect locations and progress toward more plausible replacements for the bent line. Now here is the point. With the exception of two positions held for about one second (which come shortly after the frame shown in Figure 4), none of his trial placements approximate either of the two answers that are the only solutions any student will find.

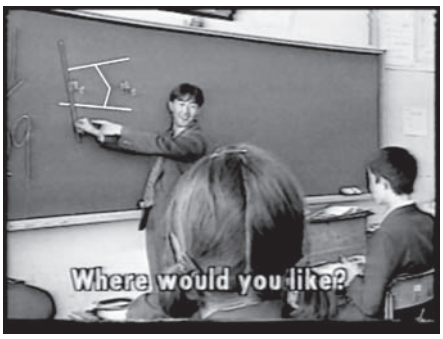



Figure 3

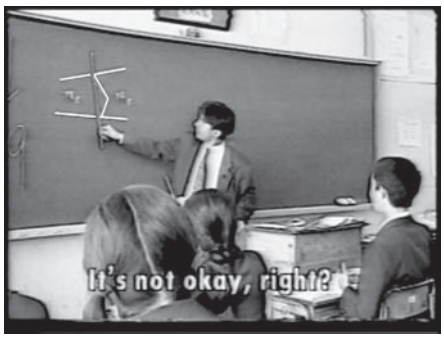



Figure 4

Rather, they are all suggestive of the orientation for the auxiliary lines that must be drawn before the basic method can be applied. He is giving subtle hints, and calling the students' attention to the very geometric features that must be noticed if the problem is to be solved. It is surely no accident that the teacher pauses with his pointer placed in two particular locations far longer than anywhere else. One of the locations is shown in Figure 4. The other is parallel to this placement, but located at the opposing vertex, which forms the bend in the boundary between Eda and Azusa.

Only after this telling warm-up – the heads-up review of the solution technique necessary to get the answer, and the casual discussion loaded with visual cues about what must be done – are the children allowed to tackle the problem.

But this is not the end of the lesson, and the students only get an announced and enforced three minutes to work individually in search of a solution.

As the children work, the teacher circulates among the students and gives hints, typically in the form of leading questions such as: "Would you make this the base? [The question is] that somewhere there are parallel lines, okay [36, p. 140]?"

He then allocates an additional 3 minutes where those who have figured out the solution discuss it with the other teacher. Weaker students are allowed to work in groups or to use previously prepared hint cards. The excerpt does not show what happens next. The TIMSS documentation [36] reports that students prepare explanations on the board (9 minutes).

Then a student presents his solution. The construction is clearly correct, and he starts out with a correct explanation. But when the time comes to demonstrate the solution, he gets lost and cannot see how to apply the area preserving transformation that solves the problem. The teacher then tells him to use "the red triangle" as the target destination.


Figure 5

The advice turns out to be insufficient, and the teacher *steps in* to redraw the triangle that solves the problem, and calls the student's attention to it with the words, "over here, over here." The student seems to understand and begins the explanation afresh. But he soon winds up saying, "Well I don't know what I am saying, but . . . " He then regains his confidence, and the presentation comes to an end without additional explanation.

A number of students say that they do not understand.

Then another student explains her answer, but the presentation is omitted from the tape. According to the Moderator's Guide [36, pp. 139–41], these two student presentations take altogether less than three minutes. Next, the teacher explains how to solve the problem. There are two equivalent answers that correspond to moving the middle vertex in Figure 1 to the left or right. Both directions solve the problem, and he shows this.

For completeness, we also show the two ways that the triangle transformation technique can be used to solve the problem. In order to make the connection between

the review material and the challenge problem absolutely clear, the problem and its two answers have been rotated to present the same perspective as the triangle transformations in Figure 1, which began the day's lesson.

Evidently, no one devised an alternative solution method.

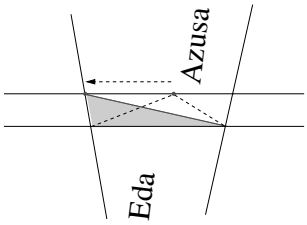



Figure 6

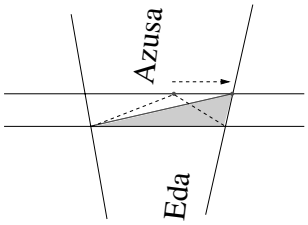



Figure 7

In his discussion of the solution, the teacher points out [36, p. 141] that this line straightening technique eliminates one of the two corners at the base of the triangle in Figures 6 and 7. This observation exposes a subtlety in that the corner that is eliminated is not the apex of the triangle, which is the point being moved to straighten out the line.

The lesson then continues with the teacher posing a new problem that can be solved with the same technique. This time the figure is a quadrilateral, and the exercise is to transform it into a triangle with the same area. At this point, the basic solution method should be within a student's reach, although the problem still requires a sound understanding of the basic method. There is also added difficulty due to the need to recognize that two consecutive sides of the quadrilateral should be viewed as representing the bent line of Figure 2, and that the other two sides should be extended as auxiliary lines to recast this new problem into a version of the Eda–Azusa exercise. The basic line straightening method can be applied so that any one of the four vertices can serve as the point where the line bends, and this designated vertex can be shifted in either of two directions to merge one of its two connecting sides with one of the auxiliary lines. The students again work individually for three minutes, and then are allowed to work in groups, use hint cards or ask the teacher.


Figure 8

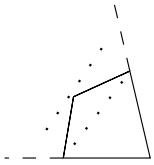


Figure 9

The TIMSS documentation indicates that this joint phase lasts for 20 minutes, and includes student presentations of their answers. There are apparently eight such presentations, which were selected to illustrate all eight ways the basic method can be applied: there are four vertices that can each be moved two ways. Then the teacher analyzes these eight ways in greater depth,

and explains how they all use the same idea. All students remain seated during this portion of the lesson, and he controls the discussion very carefully and does almost all of the speaking. For homework, the teacher asks the students to transform a five-sided polygon[6] into a triangle with the same area.

**An analysis of the teaching and its content.** This lesson is nothing less than a masterpiece of teaching, and the management of classroom time is remarkable. Although many students did not solve the first problem of the day, the assignment certainly succeeded in engaging the attention of everyone. The second problem was no giveaway, but it gave students the chance to walk in the teacher's footsteps by applying the same ideas to turn a quadrilateral into a triangle. The teacher-led study of all possible solutions masked direct instruction and reinforcement practice in an interesting and enlightening problem space.

Evidently, no student ever developed a new mathematical method or principle that differed from the technique introduced at the beginning of the lesson. Altogether, the teacher showed how to apply the method 10 times. Yet the lesson is an excellent example of how to teach problem solving, because each successive problem required a complete understanding of the basic proof technique.

The homework assignment is yet another application of the same method, and gives everyone a chance to revisit the lesson of the day once more. It also hints at the use of induction.

It is also worth pointing out that this geometry lesson, which is a specific application of measure preserving transformations, has additional uses. It appears, for example, in Euclid's proof of the Pythagorean Theorem (cf. Book I Prop 47 of Euclid's Elements).[7] More advanced exercises of this type appear on national middle school mathematics competitions in China and regional high school entrance examinations in Japan. And it is not much of a stretch to suggest that measure preserving transformations lie at the heart of those mysterious changes of variables in the study of integration.

All in all, the lesson is a wonderful example of the importance of a deep understanding of fundamental mathematics.

**Algebra.** The Japanese algebra lesson begins with student-presented answers for each of the previous day's six homework problems [36, p. 114]. These activities, along with the accompanying classroom discussion are omitted from the excerpts.

Then the teacher presents a more challenging problem that uses the same basic calculation method that the students have been studying, but needs one common-sense extension. The problem is this.

[6]The problem probably should be restricted to convex figures; otherwise it includes irregular cases that are difficult to formalize. On the other hand, this concern is just a minor technicality that has no effect on the pedagogical value of the problem.

[7]In fact, the technique is central to Euclid's development of area in general, which is based on transforming any polygon into a square with the same area. And the natural extension of this problem became a question for the ages: how to square the circle.

There are two kinds of cakes for sale. They must be bought in integer multiples; you cannot buy a fraction of a cake. The most delicious cake costs 230 yen, and a less tasty one is available for 200 yen. You wish to purchase 10 cakes but only have 2,100 yen. The problem is to buy 10 cakes and have as many of the expensive cakes as possible while spending no more than 2,100 yen.

The reproduction of the six homework exercises as shown in the TIMSS Moderator's Guide [36, p. 114] confirms that the class was already experienced with the technical mechanics necessary to solve problems with inequalities. Evidently, prior lessons had also covered word problems and the translation of word problems into equations and inequalities. Indeed, the teacher introduces the problem with the remarks, "Today will be the final part of the sentence problems [36, p. 159]." Thus, it is fair to infer that the only difference between the cake problem and the material they had just reviewed is the requirement that the solution use whole numbers of cakes.

After making sure that the students understand the problem, he asks them to devise a way to solve it. They get an announced and enforced three minutes.

Next, the teacher solicits solution approaches from the students. A student volunteers that she tried all possibilities. Her approach was to try 10 cheap cakes, then 9 cheap ones and 1 expensive one, etc., until she had the best answer. However, she was unable to finish in the three minutes that the teacher allocated for the problem. The teacher emphasizes the point, and it will soon become clear that part of the lesson is to show that this unstructured approach is unsound.

He then briefly discusses another way to solve the problem. The approach, which is quite inventive, uses a notion of marginal cost. If we buy 10 of the most expensive cakes, we exceed our budget by 200 yen. Trading in an expensive cake for a cheaper cake gives a net savings of 30 yen. Evidently, seven cakes have to be traded in, which shows that the answer is three expensive cakes and seven cheaper ones. As the teacher expected [36, p. 164], no student solved the problem this way.

Then he calls on another student, who explains how she set up the problem as an inequality, solved it as an equality, and then rounded the number of expensive cakes down to the nearest lesser integer. As she explains the equation, he writes it on the board. Only a few students understand the explanation, and he asks for another explanation of the same process. In subsequent activities that are only summarized on the tape and in the Moderator's Guide, the teacher then passes out a worksheet and works through a detailed analysis of the solution for the class.

After the detailed presentation, another problem of the same type was assigned, but with larger numbers. The teacher's words are telling:

> "If you count one by one, you will be in an incredibly terrible situation. *In the same way that we just did the cake situation, set up an inequality equation by yourself and find out . . . [the answer].* Because finding the answers one by one is hard, I wonder if you see the numerous good points of setting up inequality equations . . . "

The students worked on the problem individually. After 11 minutes, the teacher went over the problem with the class. The class ended with the teacher summarizing the solution technique that constituted the lesson of the day.

The video excerpts contain no group-based problem solving in this algebra lesson, and the Moderator's Guide confirms that none of the class time included problem solving in groups.

**An analysis of the teaching and its content.** Students never developed new solution methods. In the algebra class, the students were given the opportunity to learn first-hand why ad hoc trial-and-error approaches (which are encouraged by some of the latest reform recommendations) do not work. Although the tape does not explicitly show how many students were able to solve the original cake problem in the allotted time, the student responses suggest that no more than five could have possibly succeeded. But the three minutes of struggle might well have served to make the lesson more purposeful.

From a mathematical perspective, the cake problem was designed to require a deep understanding of inequality problems and their solutions. Mathematicians would say that when we solve a problem, we find all of the answers. If the cake problem had allowed fractional purchases, and had simply required that altogether any mix of ten cakes be purchased for at most 2100 yen, then the algebraic formulation would read,

$$230x + 200(10 - x) \leq 2100,$$

where $x$ is the number of expensive cakes purchased, and $10 - x$ is the number of the inexpensive ones. The problem would also require that $x$ be non-negative, since you cannot buy negative quantities of cake. A little manipulation gives:

$$0 \leq x \leq \frac{10}{3}.$$

Now, the point is that every $x$ in this interval is a solution to the simplified problem, and every solution to the problem is in this interval. So if we want a special answer, the interval $\left[0, \frac{10}{3}\right]$ is the place to look. If we want the largest $x$, it is $\frac{10}{3}$. If we want the largest integer $x$, it is 3. And if we wanted the largest even integer, for example, we would look nowhere else but into $\left[0, \frac{10}{3}\right]$ to conclude that this answer is $x = 2$. Incidentally, a complete answer must also observe that the number of inexpensive items must be non-negative.

This problem variant is more than a matter of common sense; it exposes students to a deep understanding of solutions to inequalities and the implications of real world constraints. Moreover, the problem illustrates the idea of decomposing a complex exercise into a more basic problem whose solution can then be adapted to achieve the original objective.

Evidently, the video excerpts feature challenge problems that cover fundamental principles, techniques, and methods of systematic thought that lie at the heart of mathematics and problem solving. As such, they ought to provide experiences that

build a powerful foundation of intuition and understanding for more advanced material yet to come. As a derivative benefit, these problems are so rich they can be readily transformed into follow-up exercises for use as reinforcement problems in class and as homework.

Both lesson excerpts exemplify a multi-round teaching and reinforcement pedagogy that begins with review of the fundamental (and systematic) principle that is the key to solving the challenge problem. The review is followed by two or three rounds (when homework is counted) that feature equivalent problems, often with additional educational content. Between each round, the teacher guides the students through the solution process to open the eyes of each learner to the basic idea, and to give the students yet another chance to apply the technique by themselves and to integrate the material into their own understanding – all in an engaging style without rote or tedium.

## 4. What can be deduced about Japanese teaching?

Many publications claim that the Japanese lessons teach students to invent solutions, develop methods and discover new principles. For example, this view is expressed in the Glenn Commission report [10, p. 4], and is clearly stated in TVCS as well: "[In Japan, the] problem ... comes first [and] ... the student has ... to invent his or her own solutions [31, p. vi]." In fact, TVCS reports that the 50 Japanese lessons averaged 1.7 student-presented alternative solution methods per class [31, Figure 22, p. 55]. Yet the excerpts exhibit no signs of such activity. They contain just one student-devised solution alternative, and it failed to produce an answer.

These differences are fundamental, and they should be reconciled. Part of the problem is that students are unlikely to devise their own solutions when the time is limited, the problems are so difficult that hints are needed, and the exercises are (clearly) designed to teach the value and use of specific techniques. Students would presumably have a better chance of finding alternative solution methods for less challenging exercises. And they would have an even better chance with problems that can be solved by a variety of methods that have already been taught. Examples might include geometry problems where different basic theorems can be used, and studies of auxiliary lines where the exercises are designed so that different auxiliary lines build different structures that have already been studied. TVCS illustrates alternative solution methods with the U.S. assignment to solve $x^2 + 43x - 43 = 0$ by completing the square and by applying the quadratic formula [31, p. 97]. Of course, this problem directed students to use different methods they already knew. The example contains no hint of any discovery.

So the question remains: where are the alternative solution methods, and when do they demonstrate signs of student-discovery?

The answers are in TVCS. It presents the actual examples that were used to train the data analysts who counted the "Student Generated Alternative Solution Methods"

(SGSM1, SGSM2, . . . ) in each lesson. The training lessons, it turns out, were the Japanese excerpts that we have just analyzed. The two student presentations for the Eda–Azusa problem are coded as SGSM1 and SGSM2 [31, p. 26–27]. Similarly, the second problem, where each of four vertices could be moved in two directions, has the codings SGSM1–SGSM8. *Altogether, this lesson is counted as having 10 student-generated alternative solution methods, even though it contains no student-discovered methods whatsoever*. And the failed try-all-possibilities approach in the Japanese algebra excerpt is counted as yet another student-discovered solution method. (See also "Teacher and Students Presenting Alternative Solution Methods [36, pp. 161–163]."")

TVCS also contains a partial explanation for the source of these judgments. It reports that the data coding and interpretation procedures were developed by four doctoral students – none of whom were in mathematics programs [31, p. 24]. Moreover, TVCS states that the project's supporting mathematicians only saw coder-generated lesson tables, and were denied access to the actual tapes [31, p. 31]. It is reasonable to infer, therefore, that they did not participate in the design of these coding practices. As for the question of invention, TVCS explains: "When seat-work is followed by students sharing alternative solution methods, this generally indicates that students were to invent their own solutions to the problem [31, p. 100]." Altogether, there appears to have been a sequence of misinterpretations that counted student presentations as alternative solution methods, which became student-generated, and then invented and which ultimately evolved into invented discoveries that might even depend on new principles the students had not yet learned ([31], [25], [10]).

On the other hand, the contributions by the Japanese teachers received much less generous recognition. Yet in the defining examples of student discovery, the teachers – not the students – manage the ideas and lead the education process.

**Additional statistics from the TIMSS projects.** It is worth reiterating that in the sample Japanese lessons, students began working individually – and not in groups – on each of the four representative exercises. Similarly, the Stigler–Hiebert analysis [30, p. 79] states that "Students rarely work in small groups to solve problems until they have worked first by themselves." TVCS contains no comparable statement, and even implies otherwise: "[After the problem is posed, the Japanese] students are then asked to work on the problem . . . sometimes individually and sometimes in groups [31, p. 134]." However, not one of the 86 figures and bar charts documents instances where problems began with students working in groups. Chart 41 [31, p. 78] indicates that of the seat-work time spent on problem solving, 67.2% of the time comprised individual effort and 32.8% of the time was spent in group-work.

Another TIMSS study addressed this issue in the statistics it gathered for a carefully balanced sampling of 3750 or so eighth graders from each participating country. One of its questionnaires asked teachers about their classroom organization and whether most of their lessons included students working in small groups, individually, as a class, etc. The results, which were weighted by the number of students

in each responding teacher's class, are reproduced below for the U.S. and Japan [3, pp. 154–155].

| Country | Percent of Students Whose Teachers Report Using Each Organizational Approach "Most or Every Lesson" | | | | | |
|---|---|---|---|---|---|---|
| | Work Together as a Class with Students Responding to One Another | Work Together as a Class with Teacher Teaching the Whole Class | Work Individually with Assistance from Teacher | Work Individually without Assistance from Teacher | Work in Pairs or Small Groups with Assistance from Teacher | Work in Pairs or Small Groups without Assistance from Teacher |
| Japan | 22 | 78 | 27 | 15 | 7 | 1 |
| United States | r 22 | r 49 | r 50 | r 19 | r 26 | r 12 |

An "r" indicates teacher response data available for 70–84% of students.

Figure 10

The table shows that Japanese lessons do not have significant numbers of small-group activities. In fact, American classes evidently contain about 4 times as many such lessons. Of course, it should be noted that the data is based on questionnaires and depends, therefore, on the judgment of each respondent. The meaning of "most or every lesson" might have cultural biases, as might the definitions of "small groups" and "teacher assistance." Still, these TIMSS statistics support the notion that the Japanese style of teaching is substantially different from many of the U.S. reform practices.

**Placing Japanese teaching in the context of U.S. reform.** The video excerpts show Japanese lessons with a far richer content than the corresponding offerings from the U.S. and Germany. TVCS reports that the eighth-grade lessons recorded in Japan, Germany, and the U.S. covered material at the respective grade levels 9.1, 8.7, and 7.4 by international standards [31, p. 44]. We suspect that the interactive nature of the teaching style, the coherent, concept-based exercises with disguised reinforcement problems, the motivated direct instruction, and the deep understanding of the teachers all contribute to the quality of the Japanese curriculum.

Additional analysis shows that 53% of the Japanese lessons used proof-based reasoning, whereas the comparable statistic for the US lessons – which included both traditional and reform programs – stood at zero [31, p. vii]. And comparisons evaluating the development of concepts – including their depth and applicability – and the overall coherence of the material likewise judged the Japanese programs to be vastly superior [30, p. 59]. By all evidence, the use of proof-based reasoning as reported in Japan is not at all representative of the reform programs in the U.S., and the use of such remarkably challenging problems is beyond the scope of any American program past or present.

When comparing U.S. reform practices and Japanese teaching methods, TVCS offers somewhat guarded conclusions that are sometimes difficult to interpret:

> "Japanese teachers, in certain respects, come closer to implementing the spirit of current ideas advanced by U.S. reformers than do U.S. teachers. For example, Japanese lessons include high-level mathematics, a clear

> focus on thinking and problem solving, and an emphasis on students deriving alternative solution methods and explaining their thinking. In other respects, though, Japanese lessons do not follow such reform guidelines. They include more lecturing and demonstration than even the more traditional U.S. lessons [a practice frowned upon by reformers], and [contrary to specific recommendations made in the NCTM Professional Standards for Teaching Mathematics] [8] we never observed calculators being used in a Japanese classroom [31, p. vii]."

Subsequent elaboration on the similarities between U.S. reform and Japanese pedagogy recapitulates these ideas in the context of various reform goals, but again offers no statistical evidence to compare with the data accumulated from the analysis of Japanese teaching practices [31, pp. 122–124]. Consequently, it is difficult – absent additional context – to compare these reform notions in terms of mathematical coherence, depth, international grade level, or the preparation of students for more advanced studies and challenging problems. And no matter what "the spirit of current reform ideas" may mean, it is clear that Japanese and U.S. reform pedagogies differ in their management of classroom time, their use of proof-based reasoning, their tradeoffs between student-discovery and the use of direct instruction, as well as their use of individual and small group activities.

For completeness, we note that TVCS makes a distinction between the idealized goals as prescribed in the NCTM Professional Standards for Teaching Mathematics, and as embodied in actual classroom practices of some reform programs. In particular, TVCS discusses two reform-style lessons. One involved students playing a game that was purported by the teacher as being NCTM compliant, but happens to have very little mathematics content: "It is clear to us that the features this teacher uses to define high quality instruction can occur in the absence of deep mathematical engagement on the part of the students [31, p. 129]." The other lesson was deemed compliant with the spirit of NCTM reforms. It began with the teacher whirling an airplane around on a string. The eighth graders then spent the period working in supervised groups to determine the speed of the plane, and came to realize that the key issues were the number of revolutions per second, and the circumference of the plane's circular trajectory. The problem also required a realization that units conversions would be needed to state the speed in miles per hour. The problem engaged the class, and a variant to compute the speed of a bird sitting on the midpoint of the string was evidently a challenge. The homework for this math class was a writing assignment: *the students were asked to describe the problem, to summarize their group's approach, and to write about the role they played in the group's work* [31, p. 127]. TVCS did not evaluate this lesson or the homework in terms of international grade level or its coherence within a curriculum.

[8] The bracketed additions are elaborations from page 123 of TVCS, where the discussion of calculator usage is reworded and thereby avoids the slight grammatical misconstruction we have caused with the unedited in-place insertion.

**Other characterizations of Japanese classroom practices.** Studies that use human interaction as a primary source of data must rely on large numbers of interpretations to transform raw, complex, occasionally ambiguous, and even seemingly inconsistent behavior into meaningful evidence. Given the complexity of the lessons, it is not surprising that different interpretations should arise. TVCS – to its credit – documents an overview of these decision-making procedures, although the actual applications were far too numerous to publish. Moreover, TVCS actually contains widely diverse observations, ideas, and conclusions that sometimes get just occasional mention, and that are necessarily excluded from the Executive Summary. Understandably, this commentary is also missing – along with any supporting context – from the one-sentence to one-paragraph condensations in derivative policy papers (cf. [25], [10]). Perhaps the seventh and eighth words in the opening line of the TVCS Executive Summary explain this issue as succinctly as possible: "preliminary findings [31, p. v]." It is now appropriate to explore these larger-picture observations and to place them within the context of actual lessons.

TVCS even offers some support for our own observations:

> "[Japanese] students are given support and direction through the class discussion of the problem when it is posed (figure 50), through the summary explanations by the teacher (figure 47) after methods have been presented, through comments by the teacher that connect the current task with what students have studied in previous lessons or earlier in the same lesson (figure 80), and through the availability of a variety of mathematical materials and tools (figure 53) [31, p. 134]."

Unfortunately, these insights are located far from the referenced figures and the explanations that accompany them. The words are effectively lost among the suggestions to the contrary that dominate the report. It is also fair to suggest that the wording is too vague to offer any inkling of how powerful the "support and direction through class discussion" really was. Similarly, the value of the connections to previous lessons is left unexplored. This discussion does not even reveal whether these connections were made before students began working on the challenge problems, or after. For these questions, the video excerpts provide resounding answers: the students received masterful instruction.

The Math Content Group analyzed a representative collection of 30 classroom lesson tables. Their assessments, as sampled in TVCS, agree with our overall observations, apart from the use of hints, which were mostly omitted from the tables. These analyses are highly stylized with abstract representations for use in statistical processing and were, presumably, not intended to be a reference for the actual teaching.[9]

[9]For example, the analysis of the excerpted geometry lesson consists of a directed graph with three nodes, two links and nine attributes. The first node represents the basic principle (attribute PPD) illustrated in Figure 1. The node's link has the attributes NR (Necessary Result) and C+ (Increased Complexity). It points to a node representing the first challenge exercise. The representations were used to get a statistical sense of various

Another sentence in TVCS begins with teachers helping students, but ends with students inventing methods.

> "The teacher takes an active role in posing problems and helping students examine the advantages of different solution methods [*however, rather than elaborating on how this takes place, the sentence changes direction with the words*], but the students are expected to struggle with the mathematical problems and invent their own methods. [31, p. 136]."

This interpretation of student work as inventive discovery appears throughout TVCS. In its analysis of the excerpted Japanese geometry lesson, TVCS categorizes the teacher's review of the basic solution method (shown in Figure 1) as "APPLYING CONCEPTS IN NEW SITUATION [31, Figure 63, p. 101]," but inexplicably switches tracks to count the student applications as invented student-generated alternative solution methods. Another such instance reads, "students will struggle because they have not already acquired a procedure to solve the problem [31, p. 135]." Similarly, TVCS never explains how teachers participate in the problem solving by teaching the use of methods and by supplying hints. Its only discussion about hinting is to acknowledge the offer of previously prepared hint cards [31, pp. 26–30]. And by the time the Glenn Commission finished its brief encapsulation of student progress, even the struggle had disappeared along with proper mention of extensive teacher-based assistance.

## 5. The matter of pedagogy

Having sequenced through the Japanese lesson excerpts to determine exactly what took place in the classrooms, we now compare these applied teaching practices with current reform principles.[10] One of the most important differences between these two approaches to teaching concerns discovery-based learning. As with any idealized theory, the real issue is how well it works in practice. Discovery-based lessons can make sense – in moderation – provided suitable safeguards are in place. In particular:

- Judgments must resolve how much time is needed for students to discover the mathematics, and the necessary tradeoffs among time for guided discovery, time for additional (or deeper) lessons, and time for practice.
- There must be detection/correction mechanisms for incomplete "discoveries".
- There must be allowances for the fact that in even the best of circumstances, only a few students will succeed in discovering non-trivial mathematical principles.

The lesson excerpts reveal a teaching style that is surprising and very different from the U.S reforms – in theory and practice. In the Japanese classes, the time allotted for the first round of grappling with problems is remarkably modest. Consequently, the

broad-brush characteristics of the lessons [31, pp. 58–69].

[10] See [2] for an enlightening albeit jargon heavy exposition on the differences between the theories of learning advanced by educators and by cognitive psychologists.

remaining time is sufficient for teacher-assisted student presentations to help identify conceptual weaknesses, and for direct instruction to present new insights, as well as for follow-up problems designed to solidify understanding. Due to the time limitations and the difficulty of the more challenging problems, many students will be learning via a model of "grappling and telling." That is, most students will struggle with a tough problem in class, but not find a solution. They will then learn by being told how to solve it, and will benefit by contrasting their unsuccessful approaches against methods that work [27]. There is no question that preliminarily grappling with a problem is both motivational and educational (cf. [4, p. 11] and [27]). And discussions to understand why some approaches fail, to understand why a solution might be incomplete, and to explore alternative problem solving techniques are all sound investments of class time. However, the use of grappling and telling raises the implementation question:

Who should do the telling?

In some teaching practices, the theory of discovery-based learning is extended to include the notion of *cooperative learning*, which holds that the students should teach one another because they "understand" each other. In contrast, the TIMSS videotape and the data in Figure 10 show that Japanese teaching is by no means purely or principally based on cooperative learning. Although students get to explain their solutions, the video excerpts show that Japanese teachers are by no means passive participants. Student explanations frequently need – and get – supervision, and students can be remarkably incoherent (cf. Figure 5) even when their solutions are absolutely perfect. When all is said and done, the teachers do the teaching – and the most important telling – but in an interactive style that is highly engaging and remarkably skillful.

Stigler and Hiebert report that the lessons do not adhere to a fixed organization. Some lessons feature more direct instruction or extended demonstrations, whereas others demand that the students memorize basic facts [30, pp.48–51]. Students might even be asked to memorize a mandate to think logically [30, p. 49].

Aharoni's article on experimental math programs in Israel deserves mention in this context. In the late 1970s, Israel developed a unique and nearly unrecognizable adaptation of the 1960s New Math, which is still in use to this day. The curriculum has been controversial; Israel had placed first on the original 1964 precursor to the TIMSS exams, and had fallen to 28th place on TIMSS 1999. Of course, this small country has experienced demographic shifts and many other sources of instability, so this drop in rank is by no means proof that the curriculum has failed, but there were other concerns about the program, and the TIMSS results gave little reason to believe that all was well.

Israel was just months away from adopting the latest U.S. reform standards when circumstances led to a reconsideration and the decision to test a program based on translations of the Singapore textbooks (from English). Aharoni is participating in this experiment, and writes about his experiences with these textbooks.

He argues that teachers must have a deep knowledge of fundamental mathematics if they are to instill a sound understanding of elementary arithmetic. His first-grade

teaching uses deep insights to provide a purposeful understanding of the most basic arithmetic operations. For example, he guides first-graders through story problems designed to open their eyes to the many different ways that a single operation – such as subtraction – can be used in the modeling of problems so that all students will enter the higher grades with the intuition and core knowledge necessary to master the translation of word problems into the native language of algebra. Only time will tell if the program is successful, but if so, his observations would have implications about best practices and teacher training.

This perspective places high demands on teachers and – by extension – on schools of education. Currently, most education programs allocate modest resources for courses on mathematics content, and very few programs are prepared to offer the kind of deep applied understanding that Aharoni describes. Instead, schools of education typically emphasize courses on developmental psychology, learning theory, and related topics such as *authentic assessment*, which is a grading practice based on portfolios of student work such as a study of how ancient Greek geometry was used 2000 years ago, or on real-life applications of periodicity – as opposed to exams. Similarly, very few mathematics departments feature course offerings on deep knowledge for K-12 instruction. This problem is further compounded by the certainty that most education majors would not have attended K-12 programs where such deep understanding would have been taught.

A small, but highly respected and widely cited comparative study by Liping Ma gives additional insight into this problem. In her study, American and Chinese elementary school teachers were asked to compute $1\frac{3}{4} \div \frac{1}{2}$, and to give a physically meaningful problem where the answer is determined by this computation. In the U.S., only 43% of those questioned performed the calculation correctly, and just one of the 23 teachers provided a conceptually correct story problem. In China, all 76 teachers performed the calculation correctly, and 80% came up with correct story problems [17].

In contrast, Hiebert and Stigler came to very different conclusions about how best to foster world-class teaching. They began with the TVCS tapes and findings, and conducted new investigations into Japanese teaching traditions. Their findings are published in *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom* [30]. According to the authors, "differences" such as "teaching techniques, ... and [teaching] basic skills [versus teaching for] conceptual understanding ... paled" in comparison to the differences they observed in the culture of teaching. In their view, the Japanese tradition of life-long reflection on how to teach, and the culture of teachers sharing these ideas among each other in a continuing process of professional development was more significant than any of these other issues, which comprise the entirety of the debate over education reform in the U.S. and elsewhere. That is, they opined that the Japanese practices of ongoing collaborative- and self-improvement were even more important than the current state of the Japanese art of teaching as well as the curriculum differences reported in their book.

However, in a follow-up videotape classroom study of teaching in Australia, the

Czech Republic, Hong Kong, Japan, the Netherlands, Switzerland, and the United States, Stigler and Hiebert came to different conclusions [11]. For this study, new data coding schemes were developed to replace those used in the 1995 TVCS. Two of the findings are particularly noteworthy. First, the new study does not mention student-invented or student-discovered solution methods, and instead of reporting an average of 1.7 student-presented solution alternatives per Japanese lesson, the new study reports that 17% of the Japanese problems featured presentations of alternative methods [11, p. 94], and that students had a choice of methods in 31% of the lessons. Second, the study found no unifying theme to explain why the stronger countries perform so well. According to the authors:

> "A broad conclusion that can be drawn from these results is that no single method of teaching eighth-grade mathematics was observed in all the relatively higher achieving countries participating in this study [12, p. 11]."

> "It was tempting for some people who were familiar with the 1995 study to draw the conclusion that the method of teaching mathematics seen in the Japanese videotapes was necessary for high achievement [11, p. 119]."

Evidently, this positional retreat (see also [11, p. 1]) must include Stigler, Hiebert, and the Glenn Commission, among others. And the fact that the follow-up videotape study did not report student-discovered mathematics suggests that the earlier finding of student discoveries was inaccurate.

These changes in understanding notwithstanding, the earlier TVCS and the follow-up *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom* will almost certainly outlive the more recent Hiebert–Stigler classroom study. These earlier publications continue to make must-read lists on education, and continue to inspire calls for reforms based on their findings. For example, on November 21, 2005, a New York Times editorial titled "Why the United States Should Look to Japan for Better Schools" cited the Teaching Gap book, and issued a call to reconsider

> "*how* teachers are trained and *how* they teach what they teach" (emphasis added).

Not one word was spent on the importance of *what content* is taught, and *what* a teacher should know in depth [29].

## 6. Conclusions

Mathematicians often ask what they can do to help preserve the integrity of K-12 math programs. In 1999, a letter protesting the new textbooks was signed by more than 200 leading American mathematicians and scientists and was published in the Washington Post. It had some positive results, but failed to stop the latest reforms. A similar protest in Israel was successful – but just barely. In California, protests

supported by grassroots parents organizations, mathematicians, scientists, concerned journalists, and politicians were able to secure a sound revision of the State K-12 math standards in 1997 – after more than five years of struggle.

In many countries, mathematics societies will probably be most effective by lobbying as a group and by seeking a role in the textbook adoptions and in overseeing the assessment programs. In the U.S., reform curricula have often been introduced in conjunction with new testing programs designed and even managed by the publishers of the newly adopted textbooks. This practice eliminates the opportunity to compare pre- and post-reform student achievement. And publishers seldom provide in-depth testing on the weakest aspects of their own programs.

It is also worth pointing out that program validation tests should cover an entire curriculum. Whereas achievement tests should concentrate on the most important material that can be covered in the allotted time, the testing of education programs should use sampling to achieve comprehensive coverage at a nominal marginal cost in the overall testing process. Needless to say, the oversight required for these assessment programs should be of the highest caliber.

Some tests use closely guarded questions. The secrecy allows the same questions to be used year after year to maintain consistency in the scoring. For example, one of the more widely cited validation studies relied mainly on a test that to the best of my knowledge has had only three of its questions appear in the literature. This achievement test was devised to align with the new math reforms, but is also reported to assess basic computational skills. It is given over a period of three days with the teachers retaining custody of the materials after school. So its questions are not really secret, and the administrative procedures lack safeguards to protect the integrity of the assessment program. Sometimes students were even allowed to rework questions from the previous day. Moreover, the test manufacturer does not require the test to be given with time limits, which are optional even for the testing of basic skills. The validation project reported year-by-year improvement of fourth-grade scores with the new reform program, but this progress was not matched by the scores for the more securely administered state testing of fifth graders.

In the U.S., the government-mandated No Child Left Behind (NCLB) testing (with state-determined tests) shows good progress for the majority of our states year by year, whereas the National Assessment of Educational Progress (NAEP) math testing shows that the net achievement of our twelfth graders has been unchanged nationwide for more than a decade. Something does not quite add up. The NAEP uses a mix of new and secret questions but is designed to be free of the biases that result from test-specific instruction and cramming. It is given to randomly selected schools, and the performance results are reported at the state level with additional results for subcategories based on gender and socio-economic status. Each student is given a randomly selected subset of test questions, and no performance results are released for students, schools, school districts, or education programs. Consequently, there is little incentive to teach to the test. The majority of the California achievement test questions are released and retired each year, and state law forbids the use of these

materials in classroom preparation for forthcoming tests. There are programs in place to detect cheating, but it is not possible to know how effective they are, and students can always use these questions for practice independently of their school assignments. In New York, there are no such prohibitions, and many New York City schools use old tests routinely in required after-school preparation sessions held during the six weeks prior to the State and City testing.

But although the NAEP may be our most uncompromised testing program, it is far from perfect. The test is consensus-based, with an oversight committee that has limited authority and where only about 10% of its members are mathematicians. The web-released sample questions suggest that the twelfth grade test is probably at a sixth grade level, on average. A representative question on fractions might be to compute two-thirds of 12 marbles. Evidently, the NAEP Governing Board (NAGB) has not reached a consensus about the benefits of knowing if an American high school education enables seniors to evaluate, say, $1/6 - 1/9$, much less $2\frac{1}{9} - 4\frac{1}{6}$.

To date, just one of the released algebra problems is categorized as solving a system of equations. This twelfth grade multiple choice question reads:

What number, if placed in each box below, makes both equations true?

$4 \times \square = \square$ and $3 \times \square = \square$: A) 0 B) 1 C) 2 D) 3 E) 4

A "hard" problem reads:

For what value of $x$ is $8^{12} = 16^{x}$? A) 3 B) 4 C) 8 D) 9 E) 12

Only 34% of our high school seniors found the correct answer even though calculators were available for use on this problem. The NAEP testing also asked students if they used a calculator for this question, but this data, unfortunately, does not appear to have been released on the web.

Needless to say, the TIMSS test questions and testing procedures, unlike many U.S. practices, stand out as a beacon of hope. But we must take care to ensure that all of the TIMSS analyses are well documented, are open to external review, and are as accurate as possible. And with so many challenges in the search for sound education reform, we may all be able to contribute somewhere in this complex of vital activities.

We close with the following summary assessments.

1. The undisciplined appeal to constructivist ideas has produced American programs that are more a betrayal of true constructivism than an advance of its principles. The result is an unprecedented reduction in the transmission of mathematical content.

2. The reform books and classroom curricula focus on examples, tricks, and experiments rather than fundamental mathematical principles, systematic methods, and deep understanding.

3. The justification for these "reforms" is based on mostly inaccurate interpretations of the best teaching practices in other countries. In particular, paradigmatic classroom examples from Japan have been misconstrued by researchers to suggest that students discover mathematical principles. In fact, the teacher conveys these

principles quite explicitly, albeit engagingly and through examples.

4. As a consequence of these misinterpretations, "exemplary" math lessons in the U.S. convey little content, take too much time, and can even lead to false "discoveries" of mathematical principles.

5. A proper understanding of best practices suggests that

i. teachers must be trained to understand, at a deep level, the mathematics they are teaching

ii. teachers should encourage individual work, but must ensure that important principles are conveyed in an orderly and cumulative manner.

6. Mathematicians, guided by proven programs such as those in Singapore, should be involved in determining the principles that are taught, the examples that help convey them, and the exercises that reinforce the net learning.

7. Mathematicians must play an active role in overseeing the quality of achievement tests in an effort to determine where our education programs are succeeding and where they are not.

**Acknowledgments.** The author is greatly indebted to Professor Michiko Kosaka for her tutoring on Japanese culture, and her verification of translated and mistranslated passages (cf. Figure 1) in the TIMSS Videotape Study documentation. It is also a pleasure to thank Clilly Castiglia and Kevin Feeley of the NYU Center for Advanced Technology and the Media Research Lab, who graciously provided the VHS frame processing.

Portions of this paper are adapted from *Testing Student Learning, Evaluating Teaching Effectiveness*, edited by W. M. Evers and H. J. Walberg [28].

## References

[1] Aharoni, R., What I Learned in Elementary School. *Amer. Edctr.* **29** (3) (2005), 8–13.

[2] Anderson, J., Reder, L., Simon, H., Applications and Misapplications of Cognitive Psychology to Mathematics Education. *Texas Education Review* (Summer, 2000). http://act-r.psy.cmu.edu/people/ja/misapplied.html

[3] Beaton, A. E., et al., *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA, 1996.

[4] Bransford, J. D., et al., *How People Learn: Brain, Mind, Experience and School*. National Research Council, National Academy Press, Wash., DC, 2000.

[5] Cobb, P., Yackel, E., Wood, T., A constructivist alternative to the representational view of mind in mathematics education. *JMRE* **23** (1992), 2–33.

[6] Cook, C., The Significance of the NCTM Standards to the Pathways Critical Issues in Mathematics. North Central Regional Educational Laboratory, 1995. www.ncrel.org/sdrs/areas/issues/content/ctnareas/math/ma0/htm

[7] Coxford, A. F., et al., *Contemporary mathematics in context: a unified approach*. Core-Plus, Janson Publications, Dedham, MA, 1997.

[8] Fendel, D., et al., *Interactive Mathematics Program: Integrated High School Mathematics 1,2 3, 4*. Key Curriculum Press, Emeryville, CA, 1997.

[9] Garfunkel, S., et al., *Mathematics Modeling Our World, 1,2,3,4*. COMAP, Inc., W.H. Freeman and Co., New York, NY, 1998.

[10] Glenn, J. et al., *Before It's Too Late, A report to the Nation from the National Commission on Mathematics and Science Teaching for the 21st Century*. 2000. www.ed.gov/inits/Math/glenn/report.pdf

[11] Hiebert, J., et al., *Teaching Mathematics in Seven Countries: Results From the TIMSS 1999 Video Study*. National Center for Education Statistics (NCES), Wash., DC, 2003.

[12] Hiebert, J., et al., *Highlights From the TIMSS 1999 Video Study of Eighth-Grade Mathematics Teaching*. NCES, NCES 2003-011, Wash., DC, 2003.

[13] Hong, K. T., et al., *Primary Mathematics 5A, 5B, 6A, 6B*. Curriculum Planning & Development Division, Ministry of Education, Federal Publications, Singapore, 2000.

[14] Jacobs, J., et al., *Third International Mathematics and Science Study 1999 Video Study Technical Report Volume 1: Mathematics*. NCES, Wash., DC, 2003.

[15] Kamii, C., Warrington, M. A., Teaching Fractions: Fostering Children's Own Reasoning, Developing Mathematical Reasoning in Grades K-12. In *1999 Yearbook of the NCTM* (ed. by L. Stiff and F. Curcio). NCTM, Reston, VA 1999, 82–92.

[16] Lappan, G., et al., *Connected Mathematics 6, 7, 8*. Connected Mathematics Project, Prentice Hall, Upper Saddle River, NJ, 1999.

[17] Ma, L., *Knowing and Teaching Elementary Mathematics: Teachers' Understanding of Fundamental Mathematics in China and the United States*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[18] Martin, M. O., et al., *School Contexts for Learning and Instruction IEA's Third International Mathematics and Science Study*. IEA, Chestnut Hill, MA, 1999.

[19] Mullis, I. V. S., et al., *TIMSS 1999 International Mathematics Report Findings from the IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. IEA, Boston College, Chestnut Hill, MA, 2000.

[20] NCTM, *Curriculum and evaluation standards for school mathematics*. NCTM, Reston, VA, 1989.

[21] NCTM, *Professional Standards for Teaching Mathematics*. NCTM, Reston, VA, 1991.

[22] NCTM, *Assessment Standards for School Mathematics*. NCTM, Reston, VA, 1995.

[23] Ocken, S., personal communication.

[24] Public Broadcasting Service, *Stressed to the Breaking Point*. High School Math Project – Focus on Algebra, TV show, MathLine series, J. Peters, Producer, 1997. See also the lesson plan at www.pbs.org/teachersource/mathline/lessonplans/pdf/hsmp/stressedtobreaking.pdf.

[25] Peak, L., et al., *Pursuing Excellence: A Study of U.S. Eighth-Grade Mathematics and Science Teaching, Learning, Curriculum, and Achievement in International Context*. NCES, NCES 97-198, Wash., DC, 1996.

[26] Ralston, A., Let's Abolish Pencil-and-Paper Arithmetic. *Journal of Computers in Mathematics and Science Teaching* **18** (2) (1999), 173–194.

[27] Schwartz, D. L., Bransford, J. D., A time for telling. *Cognition and Instruction* **16** (4) (1998), 475–522.

[28] Siegel, A., Telling Lessons from the TIMSS Videotape. In *Testing Student Learning, Evaluating Teaching Effectiveness* (ed. by W. M. Evers and H. J. Walberg), Hoover Press, Stanford, CA, 2004, 161–194.

[29] Staples, B., Why the United States Should Look to Japan for Better Schools, Editorial Observer, *New York Times*, Section A, Page 22, Column 1, Nov. 21, 2005.

[30] Stigler, J. W., Hiebert, J., *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom*. Free Press, New York, NY, 1999.

[31] Stigler, J. W., et al., *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States*. NCES, NCES 99-074, Wash., DC, 1999.

[32] Stotsky, S., ed., *What's at Stake in the K-12 Standards Wars A Primer for Educational Policy Makers*. Peter Lang, New York, NY, 2000.

[33] TERC, *Investigations in Number, Data, and Space*. K-5 book series, Dale Seymour Publications, Parsippany, NJ, 1998.

[34] U.S. Dept. of Education, *Eighth-Grade Mathematics Lessons: United States, Japan, and Germany*. Video Tape, U.S. Dept. of Education, NCES, Wash., DC, 1997.

[35] U.S. Dept. of Education, *Eighth-Grade Mathematics Lessons: United States, Japan, and Germany*. CD ROM, U.S. Dept. of Education, NCES, Wash., DC, 1998.

[36] U.S. Dept. of Education, *Moderator's Guide to Eighth-Grade Mathematics Lessons: United States, Japan, and Germany*. U.S. Dept. of Education, Wash., DC, 1997.

[37] Usiskin, Z., et al., *Geometry*. University of Chicago School mathematics Project, Prentice Hall, Upper Saddle River, NJ, 1999.

Courant Institute, NYU, 251 Mercer St., New York, NY 10012, U.S.A.
E-mail: siegel@cims.nyu.edu

# Mathematics, the media, and the public

Ian Stewart

**Abstract.** It is becoming increasingly necessary, and important, for mathematicians to engage with the general public. Our subject is widely misunderstood, and its vital role in today's society goes mostly unobserved. Most people are unaware that any mathematics exists beyond what they did at school. So our prime objective must be to make people aware that new mathematics is constantly being invented, and that the applications of mathematics are essential in a technological world. The mass media can play a significant role in encouraging such understanding, but the world of the media is very different from the academic world. I describe what it is like to engage with the media, concentrating on my own experiences of the past 40 years.

**Mathematics Subject Classification (2000).** Primary 00A06; Secondary 00A08.

**Keywords.** Popularisation of mathematics.

## 1. Introduction

For most of the 20th Century, mathematicians were free to pursue their subject essentially independently of the rest of human society and culture. In his celebrated book *A Mathematician's Apology* (Hardy [3]) the analyst G. H. Hardy wrote: 'It is a melancholy experience for a professional mathematician to find himself writing about mathematics.' In Hardy's view, writing about existing mathematics paled into insignificance when compared to creating new mathematics. In many ways he was, and still is, right. But the two activities are not mutually exclusive. Moreover, as the 20th Century has given way to the 21st, it has become increasingly vital for mathematicians to take steps to increase public awareness of their motives, activities, concerns, and contributions. Such awareness has direct benefits for the mathematical enterprise, even if that is viewed entirely selfishly: ultimately, the public purse funds our private obsessions, and will cease to do so unless the guardians of that purse are assured that the money would not be better spent elsewhere. Public awareness of mathematics (within the broader context of the 'public understanding of science') also benefits the populace at large, because we live in an increasingly technological world that cannot function effectively without substantial input from mathematics and mathematicians.

However, the role of mathematics in maintaining society is seldom appreciated – mostly because it takes place behind the scenes. The computer industry has made sure that it takes the credit (and sometimes the blame) for anything even vaguely related to its machines, but we mathematicians have failed completely to make it known

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

that without our contributions, such as algorithms (and of course much else that has nothing to do with us) those machines would be unable to add 1 and 1 and make 10. So we have a lot of work to do if we want to demonstrate that mathematics is not – as many imagine – a subject that has been rendered obsolete by the computer, but a vital part of what makes computers work. And almost everything else. To quote the preface of my recent book [9]:

> No longer do mathematicians believe that they owe the world an apology. And many are now convinced that writing about mathematics is at least as valuable as writing mathematics – by which Hardy meant new mathematics, new research, new theorems. In fact, many of us feel that it is pointless for mathematicians to invent new theorems unless the public gets to hear of them. Not the details, of course, but the general nature of the enterprise. In particular, that new mathematics is constantly being created, and what it is used for.

At the end of the 19th Century, it was not unusual for the leading mathematicians of the day to engage with the public. Felix Klein and Henri Poincaré both wrote popular books. David Hilbert gave a radio broadcast on the future of mathematics. But within a few decades, the attitude typified by Hardy seems to have taken over. Fortunately, we are now reverting to the attitudes of the late 19th Century. Distaste for mere vulgarisation gave way to grudging acceptance of its occasional necessity, and this in turn has given way to active encouragement and approval. Even today, the role of populariser is not all sweetness and light, but the days when (as happened to a colleague at another institution) a senior member of the administration would castigate a member of his academic staff for daring to write a column in a major daily newspaper are long gone. If anything, we are now more likely to be castigated for *not* not writing a column in a major daily newspaper.

Since my first appointment at the University of Warwick in 1969, indeed even before that, I have been involved in many different forms of mathematical popularisation – mainly books, magazines, newspapers, radio, and television. I generally feel much more comfortable *doing* popularisation than talking about it – in fact the main advice I give to people who are interested in becoming a populariser is to get on with it – so my intention here is to describe what it is like to be engaged in such activities, with specific reference to my own experiences. I hope that this may prove useful for others who may wish to play the role of media mathematician, and informative for those who prefer to watch from the sidelines but would like to understand the nature of the game better.

## 2. What is popularisation?

I have given many talks that popularised mathematics, but I once gave a talk *about* popularising mathematics, which is not the same thing. One example I mentioned

was a description of the Galois group of the quintic equation in comic book form [8]. Here a character in the story juggled five turnips (the 'roots' of the equation) in a blur, showing that they were indistinguishable – in short, the Galois group of the general quintic equation is the symmetric group $\mathbf{S}_5$.

A mathematics teacher in the audience objected that this was not popularisation. Just as Monsieur Jourdain, in Moliere's *The Bourgeois Gentleman*, was astonished to discover that he had been speaking prose all his life, I was astonished to discover that I had *not* been speaking popularisation. The teacher then explained that popularising mathematics meant making it accessible to children and getting them excited about it.

No: that's education. Not, perhaps, education in the sense currently envisaged in the UK (and increasingly everywhere else), which is a sterile process in which boxes are ticked to indicate that the child has temporarily mastered some small item of knowledge or technique, regardless of context, but education in the sense it used to mean, which was teaching things to children. Explaining things in a comprehensible manner, and enthusing children about the topic, are essential features of education at school level – and, indeed, in adult education too.

It was particularly clear that the teacher's view of what constituted 'mathematics' differed from mine. She was referring to the nuts and bolts of the school syllabus; my main concern was, and always will be, the frontiers of past of present mathematical research. The two are about as similar as do-re-mi and Wagner's Ring Cycle.

There is, of course, common ground. It is possible to popularise school mathematics among children without trying to teach it to them. But one of the biggest misconceptions among otherwise intelligent adults is that the 'mathematics' they did at school is *all there is*. One of the most important aspects of popularisation is to make it clear to both children and adults that this presumption is wrong.

By 'popularisation' I mean attempts to convey significant ideas from or about mathematics to intelligent, mostly sympathetic non-specialists, in a manner that avoids scaring them silly and exploits whatever interests them. I say 'attempts' because success can be elusive. The level of exposition can range from humorous short puzzles to books on hot research topics.

## 3. The public

The phrase 'public understanding of science' is widely used but seldom clarified. *Which* public? What are they supposed to understand? Why don't they understand it already?

The schoolteacher mentioned above had a very different idea of what the words 'public' and 'understand' meant, compared to what I meant. Many scientists consider the public to be anyone who is not a scientist, and view their alleged lack of understanding as a deficiency to be remedied by supplying the required information. Thus members of the public who are concerned about possible effects of genetically modified organisms are directed, by such scientists, towards research that demonstrates the

(alleged) safety of GMOs as food; people concerned about the safety of nuclear power are directed to statistical analyses of the probability of accidents, and so on. In this view, the public – whoever they may be – are considered ignorant, and the objective of the 'public understanding of science' is to remedy this deplorable deficiency.

I don't find this view helpful. Even when correct, it is patronising and self-defeating. But mostly it is not correct. Often the public, for all their ignorance of technical details, have a much clearer grasp of overall issues than specialist scientists. A major problem with GMOs, for instance, is not their safety as food, but potential damage done to the ecosystem by introducing alien species. You don't need to know any genetics to observe that numerous confident pronouncements about GMOs made by scientific experts have turned out to be wrong, and badly so. Not long ago people in the UK were assured that genetically modified DNA could not be transferred more than a few metres by pollen. It quickly transpired that such transfer routinely occurred over distances of several kilometres. It is not necessary to prove that such transfer is harmful to notice that the experts did not have a clue what they were talking about, or that their alleged expertise had led them to wildly inaccurate conclusions. On many issues of public concern, reassurance by scientists serves only to educate the public in the limitations of reductionism and the narrow-mindedness of many scientific experts.

Some scientists even seem to think that it is possible to draw up some list of basic scientific 'facts' that members of the public should know, and then teach them. So they should know that the Earth goes round the Sun, that genetic information is encoded in DNA, that the Earth is 4.5 billion years old, and so on. It would certainly help if most people were aware of such things, but this attitude encourages the view that the task of science is to establish 'the facts', and that once these are known, that's all there is to be said. Or, as a friend of mine's Head of Department put it many years ago: 'Our task as educators is to give the students the facts, and their job is to give them back to us in the exams.' Whatever that process might be, it's not education, and it's not public understanding either. Though it does help to train a lot of 'experts' who think that their limited understanding of laboratory genetics qualifies them to pronounce on the effects of GMOs on the ecology.

My view, for what it's worth, goes something like this. Let me phrase it in the context of mathematics, for definiteness: much the same goes for other areas of science.

All over the globe, every day of the week, mathematicians are carrying out research, proving new theorems, inventing new definitions, solving problems, posing new ones. The vast majority of the public have no idea that any of this is happening. They got excited by the TV programme on Andrew Wiles and Fermat's Last Theorem, but that wasn't because they thought it was the most interesting new idea in mathematics. They thought it was the *only* new idea in mathematics. What excited them was not a new breakthrough on an old problem, but the belief that for the first time in several centuries a new piece of mathematics had been brought into existence.

So the primary objective, for the public understanding of mathematics, has to be to make people aware that new mathematics is constantly being created.

This objective is more important than explaining what that new mathematics consists of, and it is more important than explaining what mathematics actually is. Only when people recognise that mathematicians are doing *something* do they start to get interested in *what* they are doing. Only when they've seen examples of what mathematicians are doing do they start to wonder what mathematics is.

If by 'the public' we really do mean the typical, randomly chosen person on the street, then we have succeeded in improving their understanding of our subject as soon as they realise that there is more to it than they met at school.

There is a more restricted subset of the public that requires, and should be given, more. These are the people who are actually interested in mathematics. They are the core audience for popularisation. For them, it is worth trying to convey more than the existence of new mathematics. It is possible to try to give a feeling for what it is.

When you watch a football match on television, it is assumed that you enjoy football and have some general idea of the rules. The commentators do not explain that the round object is a 'ball' and that the aim is to get it into the net; nor do they point out that you have to choose the right net, and that the total number of 'goals' determines who wins. You are expected to know this. On the other hand, you are not expected to know the latest version of the offside rules. The commentators assume you are aware of the issue, but have temporarily forgotten the details. By reminding you of those, they can then engage your attention in a discussion of the issues.

Too often, the media treat science very differently. You want to tell people about Fermat's Last Theorem, but first you are obliged (so the producer or editor insists) to explain what a square and a cube is and who Pythagoras was. If you want to describe the latest work on polynomial-time algorithms for primality testing, you have to explain what a prime number is and what a polynomial is. In that case, the missing information can be sketched quite quickly, but it's all too easy to find yourself in a situation where the main point you are trying to address is Galois theory, but all the programme manages to tackle is the concept of a square root. Better than nothing. . . but not what you intended, and not what is needed to break the mental link 'mathematics = school'.

## 4. Be warned

If you want to promote mathematical awareness among the public by making use of the mass media, you should be aware that it is not quite like standing in front of a blackboard or data projection screen and delivering a lecture to undergraduates. Rather different talents are needed, and in particular you have to be prepared to risk making a fool of yourself. I have dressed up in a white lab coat to talk about the probability theory of Friday 13th, presumably because the TV company concerned thought that was what mathematicians wear – or more likely thought that viewers thought that was what mathematicians wear. I have had my name up in lights on the scoreboard at Wembley football stadium, for a programme about crowd modelling

that should have taken an hour to film and actually took five because the stadium – which was supposed to be empty – was full of schoolkids on their Easter break, and was being dismantled around our ears as well.

I have spent a day lugging a stuffed duck-billed platypus round an ancient castle… a colleague, who often does TV biology, remarked 'I've never done that.' Pause. 'Mine was a stuffed echidna.' (We contemplated forming the Monotreme-Luggers' Society.) I have sat in the hot sun, visibly becoming more and more sunburned as the filming progressed, because the topic was Maxwell's equations and the backdrop of an array of radio telescopes was deemed essential to the atmosphere. I have stood in a huge supermarket at peak period to deliver five seconds of wisdom about the National Lottery to BBC news, live…terrified that the woman who was noisily changing thousands of coins at a nearby machine would still be doing so when we went on air. I have spent 16 hours in a muddy quarry filming the end of the world, and driving a battered VW beetle painted to *resemble* the world. Appropriately, its clutch-cable broke ten minutes into the filming, and I had to drive it all over the quarry, and a local farm, by crashing the gears.

One attempt at a live broadcast for Irish local radio, about alien life forms, failed because they lost the connection. We did it again the next week. In another attempt at a live broadcast – I forget what about – I sat in a cramped studio for an hour, and my slot was then pre-empted by a news flash, so nothing went out at all.

On the other hand, working with the media is occasionally wonderful. My most memorable moment ever was when we started a televised lecture by bringing a live tiger into the lecture room. (Warning: do not attempt this at home.) It's a long story, but here are the bare bones.

In 1826 Michael Faraday inaugurated a series of lectures on science for young people at the Royal Institution in London, where he was resident scientist. They have continued annually ever since, except for four years during World War II, and for almost 40 years they have been televised. Until recently they were recorded 'as live', meaning that most mistakes were left in, in front of an audience consisting mainly of schoolchildren. (Three things you should never do in show business: work with children, work with animals, work without a script. Christmas lecturers have to do all three simultaneously.) There are a few parents too, but they are placed out of sight of the cameras.

Twice in the ensuing 180 years the topic has been mathematics. Christopher Zeeman delivered the first such series in 1978, and I gave the second in 1997. One of my lectures was on symmetry and pattern-formation. We decided to open the lecture with Blake's 'Tyger' poem ('dare frame they fearful symmetry'), which, although being a cliché, seemed unavoidable.

Which meant, by the very direct logic of television, that we had to have a tiger.

A month-long search yielded a baby puma, but no tiger. We had just about decided to go with the puma when my colleague Jack Cohen found us a tiger. More accurately, a six-month old tigress called Nikka. She was wonderful, a real pro – used to the lights and an audience. She had the requisite stripes (pattern-formation, remember?). For

Health and Safety reasons she was separated from the audience by a row of upturned seats, while two burly young men held her on a chain. Apparently Health and Safety did not extend to presenters (me) so I delivered the relevant material squatting next to her. It was one of the most amazing experiences of my life, and I've never really been able to match it as a way of starting a lecture.

## 5. What the media want

When we write research papers on mathematics, the main criteria for publication are that the paper should be competently written, new, true, and interesting.

The criteria for acceptance of a newspaper article, a magazine article, a radio interview, or a TV broadcast are somewhat different. The most important difference is that you have to tell a story. A story has a beginning, an end, and a middle that joins them. Moreover, it should be clear at all times where the story is and where it is heading. This does not mean that you have to give away the punch line before you get to it: it means that the reader or listener must be made aware that a punch line is on its way. One way to describe the process is to say that the reader or listener needs to be given a 'road map', or at least a few signposts.

My feeling is that in principle even a research paper ought to tell a story, but mathematicians are not trained in narrative thinking, and readers are generally able and willing to go over a research paper several times seeking understanding. This is not the case for a newspaper article or a radio broadcast. Readers or listeners are busy people, often on their way to or from work, and they expect to be able to follow the story as it unfolds. A few may read an article twice, or record a radio programme and listen to it again, but on the whole they will do this because they *did* understand it first time, not because they did not.

As an example, suppose you want to write about Andrew Wiles's proof of Fermat's Last Theorem, for the *Ghastliegh Grange Gazette*. It does not work if you start with something like 'Let $E$ be an elliptic curve...' or even something more civilised like 'The key to proving Fermat's Last Theorem is Galois cohomology...' Instead, you need to structure the story around things the reader can readily identify with. The bare bones of the story might be something along the following lines: 'Notorious puzzle that mathematicians have failed to solve for 350 years... Very simple to state but impossible to prove until now... After seven years of solitary research, major breakthrough by Wiles... Unexpectedly linked puzzle to a different area of mathematics, making breakthrough possible... Proof temporarily collapsed after being announced... After desperate last-ditch battle, proof repaired... Triumph!'

Notice that this summary of the narrative line does not include a statement of the theorem (though you would normally work this into the article somewhere) and in fact it does not even include Fermat (though again some historical background would be a good idea). It does not mention elliptic curves or Galois cohomology, and it absolutely does not define them. Your typical reader may well be a lawyer or a greengrocer, and

these terms will be meaningless to them. If some technical idea is absolutely essential to your story, then you will have to find some way to make it comprehensible – but be aware that your readers have no idea what a function is, or a group, or even a rational number. This does not mean they are ignorant or unintelligent – after all, how much do you know about conveyancing or vegetables? It means that you are enticing them to venture into territory that is, for them, very new. They will need a lot of help. 'Infinite intelligence but zero knowledge' is a useful, though perhaps flattering, description.

Another extraordinarily important aspect of a story, for the media, is timeliness. The editor or producer will not only ask 'Why should I publish/broadcast this story?' They will ask 'Why should I publish/broadcast this story *today*?' (Or 'this week' or 'in the next available issue' or whatever.) It is not enough for the material to be important or worthy. There has to be a 'hook' upon which the story can be hung.

Typical hooks include:

- Recent announcement of the relevant research.
- Recent publication of the relevant research.
- A significiant anniversary – 100 years since a major historical figure associated with the work was born, died, or made a key discovery. A genuine professional science writer will keep a diary of such occasions, and be ready for them as they come along.
- A timely application (preferably related to stories currently considered newsworthy – such as cloning, nanotechnology, anything with a gene in it, mobile phones, computer games, the latest blockbuster movie. . .).
- A current controversy – the media always go for a dust-up, and it seldom matters if the source of the dispute is totally obscure. Everyone understands a fight.

There are other kinds of hook. With Christmas coming up, the TV programme *Esther* once decided to feature the science of Christmas, but much of it was deliberately spoof science – for example, my contribution was to point out that the aerodynamics of supersonic flight is very different from subsonic, so that at the hypersonic speeds employed by Santa Claus, reindeer antlers might be much more aerodynamic than they look.

They aren't, of course, but viewers knew it was a joke and subliminally took on board the message about supersonic flight changing the geometry. And they also got to see the back-of-the-envelope calculation that estimated Santa's speed.

One of the more bizarre hooks arose in 2003, when I received a phone call from the *Daily Telegraph*, one of the UK's major newspapers. A reader had written a letter, recalling a puzzle with 12 balls that he had heard about as a boy. All balls have the same weight, except for one. He had been told that it was possible to work out which ball, and whether it was light or heavy, in 3 weighings with a balance but no weights. Could anyone tell him how?

The response was remarkable. The newspaper reported [10] that 'By teatime yesterday [7 February 2003], *The Daily Telegraph* had received its biggest mailbag in living memory and our telephones were still ringing off the hook'. But the editors had a serious problem: it was unclear to them whether any of the proposed solutions was correct. Could I supply a definitive answer?

As it happened, Martin Golubitsky was visiting, and he remembered being inspired by this puzzle as a teenager. In fact, his success in solving it was one of the things that had made him decide to become a mathematician. We put our heads together and reconstructed one method for solving the puzzle. The newspaper duly published it, mainly as a way of ending the flow of letters and phone calls.

There is, by the way, a more elegant solution than the one we devised. The puzzle has been discussed by O'Beirne [4], who gives a solution originating with 'Blanche Descartes' (a pseudonym for Cedric A. B. Smith), see Descartes [2]. Here the hero of the narrative, known as 'F', is inspired to write the letters

F AM NOT LICKED

on the 12 balls: '... And now his mother he'll enjoin:

MA DO LIKE
ME TO FIND
FAKE COIN'

The poetic solution lists a set of weighings (four balls in each pan) whose outcome is different for all 24 possible choices of the odd ball out and its weight. The problem with this answer, clever though it may be, is to motivate it. This is why we settled for the more prosaic 'decision tree' of weighings that you will find in the published article. We felt that readers would be more likely to follow the logic, even if our method was less elegant.

## 6. The media

Let's take quick trip through the main types of media outlet. There are others – webpages, CD-ROMs, DVDs, blogs, podcasting, whatever.

**6.1. Magazines.** Popular science magazines have the advantage that their readership is self-selected for an interest in science. Surveys have shown that mathematics is very popular among such readers. Each magazine has its own level, and its own criteria for what will appeal. *Scientific American* is justly famous for the 'mathematical games' columns originated by the peerless Martin Gardner, which unfortunately no longer run.

In the UK there are *New Scientist* and *Focus*, which regularly feature mathematical items ranging from primality testing to su doku.

If you are thinking of writing an article for such a magazine, it is always better to consult the editors as soon as you have a reasonably well formulated plan. They will be able to advise you on the best approach, and will know whether your topic

has already been covered by the magazine – a problem that can sink an otherwise marvellous idea.

Expect the editors and subeditors to rewrite your material, sometimes heavily. They will generally consult you about the changes, and you can argue your case if you disagree, but you must be prepared to compromise. Despite this editorial input, the article will usually go out under your name alone. There is no way round this: that's how things are in journalism.

**6.2. Newspapers.** Few newspapers run regular features on mathematics, bar the odd puzzle column, but most 'quality' newspapers will run articles on something topical if it appeals to them. Be prepared to write 400 words on the Fields medallists with a four-hour deadline, though, if you aspire to appearing in the national news.

**6.3. Books.** Books, of course, occupy the other end of the deadline spectrum, typically taking a year or so to write and another year to appear in print. They really deserve an article in their own right, and I won't say a lot about them here, except in Section 7 below. Sometimes expediency demands a quicker production schedule. I once wrote a book in 10 weeks. It was short, mind you: 40,000 words. The quality presumably did not suffer because it was short-listed for the science book prize.

If you want to write semi-professionally, you will need an agent to negotiate contracts. At that level, book writing is much like getting a research grant. Instead of ploughing ahead with the book, you write a proposal and go for a contract with a specified advance on royalties.

**6.4. Radio.** Radio is my favourite medium for popularising mathematics. This is paradoxical, because radio seems to have all of the disadvantages (such as no pictures) and none of the advantages (such as being able to write things down and leave them in full view while you discuss them) of other media. However, it has two huge advantages: attention-span and imagination. Radio listeners (to some types of programme) are used to following a discussion for 30 minutes or longer, and they are used to encountering unfamiliar terminology. And radio has the best pictures, because each viewer constructs a mental image that suits *them*.

On radio you can say 'imagine a seven-dimensional analogue of a sphere', and they will. It may not be a good image, but they'll be happy anyway. Say the same on TV and the producer will insist that you build one in the studio for the viewers to see. TV removes choice: what you get is what they choose to show you. On radio, what you see is what you choose to imagine.

**6.5. Television.** Television is far from ideal as a medium for disseminating science, and seems to be becoming worse. As evidence: every year the Association of British Science Writers presents awards for science journalism in seven categories. In 2005 no award was made in the television category, and (Acker [1]) the judges stated:

> To say the quality of entrants was disappointing is an understatement. We were presented with 'science' programmes with virtually no science in them. Some were appalling in their failure to get across any facts or understanding. Whenever there was the possibility of unpicking a little, highly relevant, science, or research methodology, the programmes ran away to non-science territory as fast as possible, missing the whole point of the story as far as we were concerned.

I still vividly recall a TV science programme which informed viewers that Doppler radar uses sound waves to observe the speed of air in a tornado. No, it uses electromagnetic waves – the word 'radar' provides a subtle clue here. Sound waves come into the tale because that's where Doppler noticed them.

The reasons *why* television is far from ideal as a medium for disseminating science are equally disappointing. It is not the medium as such that is responsible – although it does discourage attention-spans longer than microseconds. The responsibility largely rests with the officials who commission television programmes, and the companies who make them.

Television changed dramatically in the 1980s, especially in the UK. Previously, most programmes were made 'in-house' by producers and technicians with established track records and experience. Within a very short period, nearly all programming was subcontracted out to small companies (many of them set up by those same producers and technicians) on a contract-by-contract basis. This saved television companies the expense of pensions schemes for their employees (since they now had none) and protected them against their legal responsibilities as employers (ditto). But as time passed, contracts were increasingly awarded solely on the basis of cost. A new company would get the commission to make a programme, even if they had no experience in the area, merely because they were cheaper.

Very quickly, most of the companies that knew how to make good science programmes were ousted by new kids on the block whose main qualifications were degrees in media studies and, the decisive factor, cheapness. Any lessons previously learned about how to present science on television were lost, and had to be re-learned, over and over again, by a system dedicated to the perpetual reinvention of the wheel. There is still some good TV science, but nowhere near as much as there ought to be given the proliferation of satellite and cable channels.

The good news here is that TV is once again wide open as a medium for popular science, especially now that there are hundreds of channels desperate for content. But we will have to fight all the old battles again.

## 7. Narrative imperative

Sometimes an unexpected opportunity presents itself.

*The Science of Discworld* and its sequels *The Science of Discworld 2: The Globe* and *The Science of Discworld 3: Darwin's Watch* (see [5, 6, 7]) were written jointly with Jack Cohen, a biologist, and Terry Pratchett, one of the UK's bestselling fantasy authors. They are superficially in the tradition established in *The Physics of Star Trek*, *The Science of Jurassic Park*, and *The Science of Harry Potter*, but on closer analysis they are distinctly different, and the difference is important. The latter three books all start from a popular television, film, or book series, and use that as a vehicle for *explaining* the alleged science that could actually make such things as space warps, resurrected dinosaurs, or flying broomsticks work. This approach may be an excellent way to interest non-scientists in Relativity, DNA, or anti-gravity, but it rests on a fundamental untruth: that today's science tells us that such fiction could one day become fact. But typically the true link is rather more tenuous than that between a hang-glider and an interstellar spaceship.

The Science of Discworld series takes the opposite stance. Instead of exploiting an existing body of fantasy as a basis for dubious science, it uses genuine science as a basis for new works of fantasy. In the three Science of Discworld books, we interwove entirely new fantasy stories with voyages through significant areas of modern science. We designed both aspects of the books to complement each other. The three authors worked together to plan the combined structure, choosing scientific topics that would lend themselves to a fantasy setting, tailoring the fiction to fit the facts, and selecting the facts for suitability as components of a work of fiction.

If you've not encountered Discworld before, here's a quick introduction. Pratchett's Discworld series of humorous fantasy novels now comprises 31 novels, three graphic novels, four maps, 12 plays, two television animations, a cookery book, and countless spin-offs ranging from ceramics to computer-games. Its fans are numbered in the millions. Discworld is, as its name suggests, circular in form, and flat (though decorated with forests and oceans and deserts, hills and mountains dominated by the vast heights of the central Ramtops, where the gods live in an analogue of Valhalla). The disc is about 10,000 miles across, supported by four elephants standing on the back of the great turtle A'Tuin, who swims through space.

Discworld is inhabited by people just like us, and by an assortment of wizards, witches, elves, trolls, zombies, ghosts, golems and vampires. Much of the action takes place in the city of Ankh-Morpork, where the wizards reside within the hallowed walls of Unseen University. It is a city of medieval proportions and Elizabethan filth.

Discworld was originally conceived as a vehicle for poking fun at sword-and-sorcery books, such as Robert Howard's tales of Conan the Barbarian and Fritz Leiber's 'Fafhrd and Gray Mouser' series set in the environs of Lankhmar, the model for Ankh-Morpork. But Discworld rapidly transmogrified into a vehicle for poking fun at everything, from Hollywood to the Phantom of the Opera, from religion to engineering, from the press to the police – even mathematics.

Discworld has its own sideways logic, very appealing to mathematicians. It accepts the premises of fantasy (the Tooth Fairy really does come and take away teeth, leaving real cash) but asks hard questions (what's in it for her?). It has no qualms about world-girdling turtles swimming through hard vacuum, but wonders what happens when they mate. It acknowledges the Butterfly Effect of Chaos Theory, but wonders which butterfly has this awesome power, and how to get the blighter.

Discworld is our own planet, reified. Its driving forces are magic and Narrative Imperative. In magic, things happen because people want them to. In Narrative Imperative, things happen because the power of story makes them happen. The eighth son of an eighth son cannot avoid becoming a wizard – even if the midwife made a mistake and she was actually a girl. And so, in *Equal Rites*, the misogynist wizards of Unseen University have to come to terms with a female presence in the hallowed chambers.

Some time in 1998 Cohen and I became convinced that there ought to be a book called *The Science of Discworld*. We broached the possibility to Terry, who pointed out the fatal flaw in the plan. There is *no* science in Discworld.

In vain we argued that there is. When Greebo, the hyper-macho cat belonging to the witch Nanny Ogg is shut in a box, it rapidly emerges that there are three possible states for a cat in a box: alive, dead, and absolutely bloody furious. This is a profound comment on quantum superposition: what is the association between an object's quantum state (wavefunction) and its macroscopic state (what we observe)? If you knew the cat's wavefunction, could you tell whether it was alive or dead? My own view is that you couldn't even tell that it was a cat.

Terry gently explained why this approach would be misleading. On Discworld dragons do not breathe fire because of chemistry and genetics: they breathe fire because that's what dragons do.

What saved the idea was a concept breathtaking in its simplicity. 'Terry: if there's no science in Discworld, then you must put some there.'

Thus was born the Roundworld Project, in which the wizards of Unseen University set out to split the thaum (the fundamental unit of magic) and end up coming within a whisker of destroying the whole of the universe. As the magical reactor is about to go critical and explode, taking the universe with it, the computer Hex bleeds off the excess magic to create a small sphere, a magical containment field within which magic does not work. This is Roundworld, and it is our own universe. It runs not on magic, but on rules. It has helium and magnesium, but no narrativium. Things happen there because the rules say they must, not because someone wants them to.

Oddly, this makes everything in Roundworld harder to understand, not simpler. If a person wants something built – a house, say – then they get some builders and up it goes. But if the *rules* want something built, such as a human being, then the construction process is much more obscure, involving big molecules and bacterial blobs and billions of years of nothing much happening; then blink your eye and the humans have come and gone, leaving only the ruins of the Space Elevator, and you can't even be sure they were human.

The best way to envisage the structure of the Science of Discworld books is to think of a novelette by Pratchett, set on Discworld, with its usual cast of characters and its usual narrative constraints, but with Very Big Footnotes by Cohen and Stewart. The novelette, which comprises the odd-numbered chapters of the book, is fantasy; the footnotes, comprising the even-numbered chapters, are the scientific commentary, and are typically between two and three times as long.

This is a beautiful framework for writing about science, because the differences between magic and science are highly illuminating. Discworld is the perfect framework for a 'What if?' discussion of science – a well-established, self-consistent universe that can be used to 'compare and contrast'. We managed to work quite a lot of mathematics into the books, too: chaos, complexity, Langton's Ant, probability, phase spaces, combinatorics, information theory, infinity, and transfinite numbers. Not to mention scores of applications from astronomy to zoology.

It was fun, too.

## References

[1] Acker, F., Dorks' night out. *The Science Reporter* (Nov/Dec 2005) 1–2.

[2] B. Descartes, B., The twelve coin problem. *Eureka* **13** (1950) 7, 20.

[3] Hardy, G. H., *A Mathematician's Apology*. Cambridge University Press, Cambridge 1940.

[4] O'Beirne, T. H., *Puzzles and Paradoxes*. Oxford University Press, London 1965.

[5] Pratchett, T., Stewart, I., and Cohen, J., *The Science of Discworld*. Ebury Press, London 1999.

[6] Pratchett, T., Stewart, I., and Cohen, J., *The Science of Discworld II: The Globe*. Ebury Press, London 2002.

[7] Pratchett, T., Stewart, I., and Cohen, J., *The Science of Discworld III: Darwin's Watch*. Ebury Press, London 2005.

[8] Stewart, I., *Ah, Les Beaux Groupes!* Belin, Paris 1983.

[9] Stewart, I., *Letters to a Young Mathematician*. Basic Books, Cambridge 2006.

[10] Uhlig, R., Odd ball letter starts maths puzzle mania. *Daily Telegraph*, 8 February 2003, 7.

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK
E-mail: ins@maths.warwick.ac.uk

*Panel A*

# Controversial issues in K-12 mathematical education

Michèle Artigue (*moderator*)
Ehud de Shalit and Anthony Ralston (*panelists*)

**Abstract.** This article sets the background for the panel session at the ICM on controversial issues in K-12 mathematics education. Three specific issues have been selected: Technology, skill building and the role of test and assessment. For each of these, a list of questions has been prepared. After introducing the three themes and the associated questions, this article presents the positions on these of the two panelists: Professor Anthony Ralston, from the State University of New York at Buffalo in the US, and Professor Ehud de Shalit from the Hebrew University of Jerusalem in Israel. The article ends with some personal comments from the coordinator of the panel: Professor Michèle Artigue from the University Paris 7 in France.

**Mathematics Subject Classification (2000).** 97A80, 97D30, 97D40, 97D60.

**Keywords.** Mathematics education, K-12 curriculum, technology, skill building, concept building, testing, assessment.

## Introduction

by *Michèle Artigue*

K-12 mathematics education is obviously a controversial area, so much so that, in countries like the US, the term Math Wars has been used for describing the kind of conflicts between communities that has been generated in recent years. We all regularly hear colleagues complaining that the students they receive have not been adequately trained and that, every year, the situation becomes worse, or that they are not pleased with the kind of mathematics education their children receive etc. We all know that such feelings are not something new, but we cannot deny that in the last decade they have dramatically increased in intensity in many countries.

Why does such a situation exist? What are the real challenges that K-12 mathematics education has to face at the beginning of the XXI$^{st}$ century? What can mathematicians do in order to enhance or support efficiently the necessary efforts, evolutions and changes of the whole educational community? These are the crucial issues that motivate the existence of a panel session on Controversial Issues in K-12 mathematical education at the ICM2006 in Madrid. It is certainly interesting to keep them in mind even if the panel does not address them all directly.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

For structuring this panel session, we have selected some particularly controversial issues, and will try to elaborate on these, with the support of the audience. These issues approach the current problems met by K-12 mathematics education through three different, but not independent, topics: technology, the place given to the learning of skills and techniques, and assessment and tests. Everyone will certainly agree that each of these is today a controversial topic, and that frequently in what we read or hear, it is advocated that the ways they have been dealt with in recent years or currently has resulted in some of the difficulties in K-12 mathematics education today.

In what follows, we briefly introduce these three topics and articulate some questions that we would like to discuss for each of them. We then present the positions on these questions of the two panelists, Professor Anthony Ralston from the State University of New York at Buffalo, and Professor Ehud de Shalit from the Hebrew University in Jerusalem. The article ends with some general comments by the moderator of this panel session, Professor Michèle Artigue from the University Paris 7.

## Topics and questions

**Technology.** In 1985, the first study launched by ICMI entitled "The influence of computers and informatics on mathematics and its teaching" was devoted to computers and the ways the learning and the teaching of mathematics as well as this discipline itself was affected by technology. A second edition of the book issued from this study was prepared by B. Cornu and A. Ralston and published in 1992 in the Science and Education Series of UNESCO. As described in its introduction, the UNESCO book addresses the importance of the changes introduced by technology in professional mathematical practices and makes suggestions for new curriculum elements based on these new methods of doing mathematics. It is pointed out that even if these suggestions are judged by the reader to be stimulating and even persuasive as well as reasonably grounded, it is nevertheless the case that "such suggestions are fundamentally speculative at the level of large scale implementation – by which we mean that converting them into a well-developed and tested curriculum for the typical teacher and the typical student is still a major challenge."

Since that time, more and more sophisticated technological tools have continued to be developed for supporting the learning and teaching of mathematics, and their use is today encouraged by the K-12 mathematics curriculum in most countries. Nevertheless, in spite of the existence of an increasing amount of positive small-scale experiments, the real nature of the effect of technology on mathematics education in the large remains under discussion. The problems raised in the first ICMI study have not been solved, and the discourse of those who think that the impact of technology is globally negative and ask for a strict limitation of the use of calculators and software, and even for their banishment from mathematics education in the early grades, is opposed by those who consider that it does not make sense today to think about mathematics learning and teaching without taking into account the existence

of technology and without trying to benefit from the real and increasing potential it offers for mathematics education.

Thus the first set of questions we propose to raise is:

*Up to what point should the changes introduced in social and professional mathematical practices by technology be reflected in mathematics education?*

*What does technology have to offer today to K-12 mathematics education and why does it seem so difficult to have it benefit mathematics education in the large outside experimental settings?*

*What could be done in order to improve the current situation?*

*Is a strict limitation on the use of calculators and software a reasonable solution?*

**Skill building.** Every one of us certainly agrees that mathematical learning, as with any kind of human apprenticeship, requires skill building and also that it requires much more than that. In recent decades all over the world, K-12 mathematics curriculum developers, influenced by constructivist and socio-constructivist epistemologies of learning, by the results of cognitive research on learning processes, and also by the observed limitations of students' achievements in mathematics, have stressed the necessity of moving some distance from teaching practices seen as too focused on drill and practice, and of getting a better balance between the technical and conceptual facets of mathematical learning. K-12 mathematics curricula have given increasing importance to exploration and work on rich and open problems in order to help students understand better the reasons for mathematical conceptualizations, and these conceptualizations themselves. They have also promoted teaching strategies that try to give more importance to the personal and collective elaborations of students in the development of classroom mathematical knowledge. Once more, the global effects of these curricular changes on K-12 mathematics education are a matter of controversy. Voices have arisen asking for a radical change in the role to be given to the learning and mastery of algorithms, with the long division algorithm often appearing as emblematic of the desired changes. In a similar vein, other voices denounce the dangers of what they see as a new "back to basics" program and the inability to understand that mathematics teaching has to take into account social and technological evolution, and the changes in scientific and mathematics culture needed in our societies today.

Thus a second set of questions:

*What is the pertinence of the opposition between skill learning and the exploration of rich problems? Between techniques and concepts?*

*What is the right balance to be achieved in K-12 mathematics education between the different facets of mathematical activity?*

*How can this balance be achieved and what are the respective mathematics responsibilities to be given to the teachers and the students?*

**Test and assessment.** We are all aware of the influence that the form and the content of assessment have on any form of education and, thus, on K-12 mathematics education. We are also aware of the increasing importance given to national and

international testing, as reflected for instance by the coverage in the media of the PISA enterprise of the OECD and TIMSS, and the influence that these results are taking in educational policies. The importance to be given to external assessment versus internal assessment, to international comparisons and standardized testing, to the effect of assessment on the mathematics learning of students, and to the effect of systematic testing on educational systems are all controversial issues, as are the discussions generated by the "No Child Left Behind" legislation in the US. Thus our third set of questions:

*How can we correctly reflect in assessment what we wish to achieve through mathematics education?*

*Is standardized testing ever useful? For what purpose? Under what conditions?*

*What exactly is tested by international assessments such as PISA or TIMSS? Do they represent the mathematical culture that we want K-12 mathematics education to develop? What can we learn from them?*

## A reform perspective

by *Anthony Ralston*

**Preamble.** I believe passionately that the K-12 mathematical curriculum, as it exists in most countries, needs substantial reform. But, because the notion of "reform curriculum" means different things to different people, I think I should begin by delineating the perspective from which I view the reform of mathematics curricula.

First, neither constructivism nor its antithesis plays any role in my beliefs about reform. Thus, arguments about such things as discovery learning or about whether rote memorization is a good or a bad thing will play no role in what follows here.

Next, I believe strongly that mathematics should be a demanding subject in all grades, probably the most demanding that students study in each grade. Thus, any suggestion that mathematics should be "dumbed down" at any level is anathema to me.

Finally, I believe, as surely all attendees of ICM2006 do, that mathematics is a dynamic, growing subject with ever-changing opinions on what is more important or less important mathematical subject matter. But, perhaps in contradistinction to many ICM attendees, I think this perspective must include not just areas of research but also the entire K-12 curriculum. Thus, what is important subject matter in K-12 mathematics today may be – I think, is – different from what it was yesterday and no doubt is different from what it will be tomorrow.

**Technology.** Mathematicians were slower than almost all scientists and engineers to make computing technology a part of their everyday working lives[1]. Nowadays, how-

[1]Mathematicians' attitudes about technology as well as about other matters considered in this paper are discussed in A. Ralston, Research Mathematicians and Mathematics Education: A Critique, *Notices Amer. Math. Soc.* **51** (2004), 403–411.

ever, many research mathematicians use computers routinely for number crunching, for accessing computer algebra systems, and for using a variety of other computer software for both professional and non-professional purposes. Still, it appears that, even as most mathematicians now recognize computer technology as an indispensable tool for doing mathematics research, they resist the notion that computers should be widely used in mathematics education on the grounds that what is important in K-12 mathematics education has hardly changed in – dare one say it? – the past century.

The crucial aspect of whether – and, if so, when – computers or calculators should be used in K-12 mathematics education has resulted in more controversy than any other aspect of mathematics education. I have written elsewhere about my belief that pencil-and-paper arithmetic (p-and-p, hereafter) should be abolished from the primary school curriculum in the sense that no level of proficiency in it should be expected of students although teachers should be free to use p-and-p examples as they wish. Since I published a paper to this effect in 1999[2], I have seen no reason, cogent to me, to back off from this position[3]. Of course, you must understand that, keeping in mind the position stated in the Preamble, I would replace a p-and-p-based curriculum with a rigorous curriculum emphasizing mental arithmetic while allowing free use of calculators in all grades. The goal of such a curriculum, as with any arithmetic curriculum in primary school, would be to achieve the *number sense* in students that would enable them to proceed successfully with secondary school mathematics.

I cannot provide any evidence why a mental arithmetic, calculator based curriculum would work because it has not been tried but neither has anyone adduced a compelling reason why it should not work. Moreover, no one can give good reasons to continue the classical p-and-p curriculum which has never worked very well and must now be working more poorly than ever, given that almost all students will recognize that the classical curriculum tries to teach them a skill without practical value any longer. In addition, since students will almost universally use calculators outside the classroom, forbidding them inside the classroom is self-defeating. Only if it can be argued that a p-and-p-based curriculum is clearly the best way to prepare students for subsequent study of mathematics, can such a curriculum be justified in the 21st century. But I don't believe any compelling argument of this nature can be made; all such attempts I've seen can only be described as feeble.

Learning *arithmetic* – what the operations are, when to use them, place value etc. – is crucial for the study of all subsequent mathematics. But not only is p-and-p calculation not necessary to the goal of learning about arithmetic, it is positively destructive of that goal.

[2]A. Ralston, Let's Abolish Pencil-and-Paper Arithmetic, *Journal of Computers in Mathematics and Science Teaching* **18** (1999), 173–194.

[3]An area of particular controversy is whether the traditional long division algorithm should be taught at all. My opinion on this can be found in A. Ralston, The Case Against Long Division, http://www.doc.ic.ac.uk/∼ar9/LDApaper2.html.

**Skill building.** Skill building is of value in K-12 math education only insofar as the skills learned facilitate the doing of mathematics and the subsequent study of mathematics. It must be recognized that (almost?) none of the skills traditionally taught in K-12 mathematics have value any longer as skills per se. But, following the foregoing argument, if p-and-p skills are not to be taught, it is imperative that learning substantial mental arithmetic skills should be a major goal of primary school mathematics. These skills should include not just the obvious ones of immediate recall of the addition and multiplication tables and the ability to do all one-digit arithmetic mentally but also the ability to do substantial amounts of two-digit arithmetic mentally.

It needs to be emphasized that the development of good mental arithmetic skills requires good coaching from a teacher about the various algorithms that can be used to do mental arithmetic and then hard work by the student. Mental arithmetic, say two-digit by two-digit multiplication, is hard[4]. Learning to do it well involves much practice during which the student will decide which algorithm is most congenial to her/him. Teaching and learning mental arithmetic must be a joint responsibility of teacher and student.

One advantage of learning to do two-digit arithmetic mentally is that such a skill requires a good grasp of place value, an important aspect of primary school mathematics in any case. Another advantage is that automaticity or near automaticity in one- and two-digit mental arithmetic allows students to be given demanding word problems. More generally, sound technique in mathematics must always be the forerunner of good conceptual understanding.

A word about fractions. Primary school is certainly the place where students should learn about fractions, reciprocals and the conversion of fractions to decimals and vice versa. But I doubt it is the right place for them to learn fraction arithmetic except perhaps in some simple cases. When students get to secondary school, they will need to do arithmetic on algebraic fractions. This would be the best time to teach the arithmetic of both numeric and algebraic fractions since, in any case, few students will remember the arithmetic of numeric fractions from when it may have been taught in primary school.

**Test and assessment.** The standardized testing culture that has swept over the United States and is rapidly advancing in the United Kingdom and other countries is perhaps the most serious threat of all to quality mathematics education throughout the world. The standardized testing requirements in the U. S. No Child Left Behind (NCLB) legislation will have the almost certain result that NCLB will be that act most destructive of quality education ever passed by the United States Congress.

The pressure on schools and teachers for students to achieve high grades on stan-

[4]Is there any reason why learning to perform two-digit by two-digit multiplication mentally should not be a realizable goal of school mathematics? I don't think so. Some positive evidence is contained in D. Zhang, Some Characteristics of Mathematics Education in East Asia – An Overview from China, in *Proceedings of the Seventh Southeast Asian Conference on Mathematics Education* (N. D. Tri et al., eds.), Vietnamese Mathematical Society, Hanoi, 1997.

dardized tests always leads to a number of evils that have been widely catalogued. Three of the worst are teaching to the test, emphasis on routine mathematics at the expense of advanced topics and problem solving, and the inordinate amount of time taken to prepare for these tests which not only drives important mathematics from the classroom but also often means decreased attention to science, history and the arts generally. Moreover, the inevitable result of emphasis on standardized tests is that scores increase without any concomitant increase in learning[5].

I am not opposed to testing students. Quite the contrary. It is by far the best way for a teacher to assess the learning of her/his students. But in the not quite antediluvian past, the assessment task was left to individual teachers in their classrooms. Why have things changed so much? The answer in the United States and other countries appears to be that educational administrators, politicians and even parents no longer trust classroom teachers to do the assessment job themselves. This is not altogether wrongheaded. As I and others have argued elsewhere[6], the quality of K-12 mathematics teachers in, at least, American schools has been declining for half a century and, while there are still many excellent mathematics teachers in American schools, too many are not competent to teach the mathematics they are supposed to teach[7]. But, if this is so, standardized testing will only exacerbate this problem by convincing too many who might become teachers that there is no scope for imagination or initiative in school mathematics teaching.

The crucial point is that there is no sign whatever that standardized testing has ever been effective in increasing student learning. If all standardized testing in all subjects were abandoned at all levels short of university entrance, this would be an immediate boon to all education.

I should say a word about TIMSS and PISA. Since both of these are essentially diagnostic tools given to a sampling of students, they do not suffer from most of the strictures above. For example, teachers cannot teach to the test because at most a very few students in each class will take these tests.

## A traditional perspective

by *Ehud de Shalit*

The author of this essay is a mathematician who found himself involved in questions of mathematical education despite lack of formal background in the discipline. I make

[5] See A.Ralston, The Next Disaster in American Education. *The Sacramento Bee*, 1 December 2002 (http://www/doc.ic.ac.uk/∼ar9/NextDisaster.html).

[6] See A. Ralston, The Real Scandal in American School Mathematics, *Education Week*, 27 April 2005 (http://www/doc.ic.ac.uk/∼ar9/TeacherQual.html) and V.Troen and K.C.Boles, *Who's Teaching Your Children? Why The Teacher Crisis is Worse Than You Think and What Can Be Done about It*, 2003, Yale University Press.

[7] Indeed, while mathematicians generally choose to argue about something we may be knowledgeable about – curriculum – a far more serious problem with mathematics education in most countries is the inability to attract enough high quality people to teach school mathematics.

no claim to know the literature of science education, and I am surely ignorant of important studies in the area. I nevertheless dare to participate in the discussion because I believe that educators and scientists alike should bear the burden of shaping our children's education, listening to and learning from each other's point of view. It is deplorable that recently, the two communities of math educators and mathematicians have been poised against each other, mostly, but not always, the first being portrayed as "reformers", the latter as "traditionalists"[8]. Emotions have run high, and the two communities found themselves in conflict, instead of joining forces towards a common cause.

This being said, I also want to apologize for not having equally strong opinions on all issues. In fact, I will address two of the points raised by Prof. Artigue (*the impact of technology* and *skill building*), and make only minor remarks on the third (*tests and assessment*), which I consider to be a political issue more than a mathematical or educational one. I hope to make myself clear in due time. Moreover, depending on the circumstances, these three sample topics, important as they be, need not have a decisive affect on the success or failure of a given system. External factors such as class size, discipline, teacher training and resources, which vary considerably from state to state, are often of greater importance than questions of curriculum and methodology. However, unable to influence the first in a direct way, we, mathematicians, focus on the latter.

My starting point is that mathematics teaching *need not* necessarily follow the rapid changes in the usage of the subject in society or technology. Its prime role is to imbed in our children a basic sense for, and understanding of numbers, symbols[9], shapes and other "mathematical objects", together with skills in manipulating these objects, that are needed to develop what is commonly called "mathematical reasoning". The objects to be chosen, the time devoted, and what is taught about them, should be dictated by their prominence in mathematics, and their epistemic and pedagogical value, and less so by their frequency in daily life. This does not mean, of course, that examples and applications of the material should not be updated and modernized, but I do preach respect for the traditional way of teaching, because more than it was based on old *needs*, it was based on inherent *values* that have not changed with time. A well-trained mathematical mind is a highly flexible system. If brought up correctly, it will find its way to adjust and analyze mathematical scenarios very different from the ones that surrounded it initially, while it was being shaped.

As an example, consider the well trodden issue of long-division. I believe that the standard algorithm should be taught in elementary school, thoroughly explained and practiced *not* because of its practical value. Rather, it is important because it enhances the understanding of the decimal system, of zero as a place-holder, of the Euclidean algorithm, and is a necessary precursor for polynomial arithmetic. It allows

[8] Those unaware of the ongoing controversies, can read David Ross' article *Math Wars* (www.ios.org/articles/dross_math-wars.asp) and the references therein.

[9] A. Arcavi, Symbol sense: informal sense-making in formal mathematics. *For the Learning of Mathematics* **14** (3) (1994), 24–35.

the child to review the multiplication table and develop number sense while doing something else, more advanced, so it makes learning more interesting. Moreover, it is natural. It therefore agrees with *mental arithmetic*, and helps us visualize the process involved in division. For these and for many other reasons, well explained in[10] and not mentioned here for lack of time, long division is a pedagogical gold mine. The abandoned algorithm for extracting the square root, often cited to justify abolishing long division as well, is in comparison a pedagogical swamp, was abandoned for this reason, and not because it became obsolete.

Respect for traditional values in education has another advantage, that new theories are tested gradually, and radical potentially damaging changes are avoided. An ailing educational system need not be ailing because its underlying principles or methods are old-fashioned, and *reform* in itself is not an automatic cure, even where needed. More than often, the reason for failure is that good old principles stopped being implemented correctly, for various sociological reasons on which I do not want to elaborate here.

The second general remark is that I do not believe in teaching in vacuum, or in a content-empty environment. Learning must focus on concrete concepts, methods, algorithms if necessary. Insight and creativity come with variations on a theme, not where there is no theme. Teaching "how to solve it" is not synonymous with dry cookbook mathematics. It can be fun and enlightening. Constructivism[11] has led some educators to minimize teacher's intervention in the learning process. Such an approach may be tried on a single-time basis, through enrichment activities. But it is time consuming, with the average teacher may lead to fixation of mistakes, and for anyone but the brightest students can be very frustrating. We simply cannot expect the children to come up with the great discoveries of arithmetic and geometry, let alone calculus, by pure exploration. A fundamental feature that distinguishes human beings from animals is that we can learn not only from our own experience, but also from that of our ancestors. To be illustrative, I think of the art of teaching as give-and-take. The teacher delivers a package of knowledge, bit by bit, each time taking back from the students their responses, their reflections, their mistakes. On these she or he builds up, shaping and manipulating the dialogue, until a deep understanding and the desired proficiency are achieved. To believe that these can spring up spontaneously, just by setting the stage and giving a slight stimulus, is to assume too much.

Finally, a word about the term *conceptual thinking.* It is often brought up by advocators of certain approaches in education to distinguish their goals from those of others, who – so it is to be understood – lead to lower level thinking. I don't know of any kind of thinking that is not conceptual. Abstraction, in language or in mathematics, making generalizations, or conversely, looking for examples, testing predictions and searching for the right vocabulary to communicate our mental processes, are all instances of conceptual thinking, namely thinking in terms of concepts. The contro-

[10] *The role of long division in the K-12 curriculum,* by D. Klein and J. Milgram, ftp://math.stanford.edu/pub/papers/milgram/long-division/longdivsiondone.htm.

[11] *Constructivism* is the cognitive theory based on the idea that knowledge is constructed by the learner.

versy, in my view, is not about whether conceptual thinking is more or less important than basic skills, but whether acquiring those skills is part of conceptual thinking, as I want to argue, or not[12].

**Skill building.** Drill and practice. Like a swimmer or a pianist the student of mathematics has to absorb great ideas, but also to practice hard to be able to use them efficiently. Contrary to the common belief, the primary reason for skill building is not the need to perform mathematical tasks with great precision and speed, because in our age these human qualities have been surpassed by machines, and we need not regret it.

I see three important reasons to promote skill building. The first is that skill building is essential for forming a sense for numbers, and later on for symbols, functions, or geometry. Subtle instances of insight and analogy, are woven into a web of images and associations in one's mind, and cannot be classified and taught sequentially. They are only the product of long-term practicing and skill building. The distance between knowing something in principle and mastering it is very big in mathematics.

The second reason is that our mind functions on several levels simultaneously, and we are not always aware of the sub-conscious levels that are "running in the background", if I may use a metaphor from computer science. To be able to free the thinking creative part of our mind, to let it form the web of links needed for exploration and discovery, we must defer to the background more routine tasks, that in the past occupied the front, but should now be performed semi-automatically.

The last reason in favor of skill building is rarely mentioned, and might seem to you heretical. Experience has taught me that many children, especially those suffering from math phobia or learning disabilities, are highly rewarded psychologically by success in performing a routine algorithm, such as long division, and by acquiring proficiency in a given task. Such a reward for them is a higher boost than the ability to understand the theory behind it, or the fun in discovering a method by themselves. Once they know the "how" they are lead to ask "why". I would not rule out an approach that harnesses skill building before understanding, if the teacher feels that it suits the child better. Needless to say, both aspects should eventually be covered, and bright children who have mastered the technique and eagerly ask good questions should not be hindered.

Skill building is often confined – by those promoting "conceptual understanding" as a substitute – to algorithmic skills, and algorithmic skills are then downgraded to mere rote. While algorithmic skills are very important, and the algorithmic approach to arithmetic is something to be cherished, as I made clear in the example of long division, mathematical skills are by no means only algorithmic. The ability to translate a word problem into arithmetic, or later on into algebra, is a well-defined skill. Analyzed closely, it consists of many sub-skills, like distinguishing relevant information from irrelevant data, choosing the variables cleverly, translating prose into algebra, and

[12] See *Basic skills versus conceptual understanding, a bogus dichotomy in mathematics education,* by H. Wu (http://www.aft.org/pubs-reports/american_educator/fall99/wu.pdf).

finally the technique of solving, say, a system of linear equations. Geometric skills, drawing to scale, recognizing hidden parts, decomposing and assembling figures, as well as computational skills of area and volume, form another category.

Given my earlier criticism of the constructivist approach, it will not come as a surprise that I believe in *standard algorithms*. It is true, students who come up with their own (correct!) algorithms should never be scolded, but eventually standard algorithms are more efficient, help in the process of automatization of algorithmic tasks discussed above, and also serve an important purpose of establishing a common language.

As an example, after a certain amount of preparatory classes meant to clarify the distributive law, which may include both manipulations of brackets and geometric representation by rectangles, I would simply *teach* the standard algorithm for "vertical multiplication". I do not see the benefit in letting the students make up their own algorithms, where inevitably many will multiply units with units, tens with tens etc. and then add them up. To expect from fourth graders to come up with what was one of the main achievements of the Hindus and the Arab scholars in the Middle Ages is unrealistic. However, once the algorithm has been explained, both the *how* and the *why*, and practiced, there are many subtle questions that can be left for discussion and discovery. Would it always be more economical to apply the algorithm as is, or perhaps switching the position of the two numbers to be multiplied saves some operations? How can we estimate in advance the order of magnitude to save us from potential pitfalls, what double-checks should we make etc. etc.

Anthony Ralston, in his paper "*Let's abolish pencil and paper arithmetic*"[13] advocates to abolish basic algorithmic skills that were the bread-and-butter of elementary school arithmetic for centuries. He summarizes his discussion by saying "*Since no one argues any longer that knowledge of PPA* (*pencil-and-paper arithmetic*) *is a useful skill in life* (*or, for that matter, in mathematics*)*, the question is only whether such 'deprivation' could leave students without the understanding or technique necessary to study further mathematics.*"

Even if we accept the premises, doubtful in my mind, I think he misses the point. First, any attempt to separate understanding from technique is artificial. Second, it is the miracle of the subject that the very same principles underlying higher mathematics, or fashionable topics such as geometry and statistics, often quoted as benefitting from the time freed by the abolishment of PPA, are manifested in their purest and simplest form in these basic skills. A person not knowing how to calculate what $\frac{3}{5}$ of $4\frac{2}{7}$kg of rice are, will not have the technique to analyze the changes in the school budget of England. Nor will he have developed enough intimacy with numbers to estimate those changes in advance, or tell instantly, if his calculator-based computations make sense or not.

As a substitute to PPA, Ralston elevates *mental arithmetic* to a central position in his proposed program. To give examples, he expects elementary school students to

[13] In *Journal of Computers in Mathematics and Science Teaching* **18** (2) (1999), 173–194.

perform two-digit by two-digit multiplication mentally, and high-school students "to be able to factor a variety of three term quadratics mentally". To succeed, he admits, mental arithmetic should be practiced in calculator-free environment. I wholeheartedly agree with the importance of mental mathematics, both for developing number (and symbol) sense, and for practical purposes, estimation and checks. I do not understand though the reluctance to allow one to put things on paper. PPA does not contradict mental arithmetic. It records it, something we shouldn't be ashamed of, and without which we cannot communicate or analyze peacefully what we have done. It also helps in visualizing graphically the steps carried in our mind, and it allows us to organize little mental steps into a larger procedure, without putting too heavy a burden on our memory.

**The impact of technology.** There are two somewhat separate questions here. The first is to what extent should the curriculum be dictated by the way mathematics is used in technology, and to what extent should we conform to requests coming from the changing society, rather than teach basic principles and skills[14]. I have expressed my opinion about this question in the opening statement. Contrary to the quotation just mentioned, I believe that education in the large, ought to enrich the child and teach him or her basic skills, knowledge, values and understanding that are *absolute*. If carried out correctly, they will inevitably produce a knowledgeable, thinking, skilled and creative citizen. If tailored to the needs of a certain industry or society, rather to these absolute values, they will produce poor technocrats.

The second question involved in the issue of technology is to what extent do technological innovations influence the way we teach in class. This concerns mostly calculators in elementary school, but also the use of graphic calculators in calculus, Excel sheets in statistics and computers in general.

It would be wrong to ignore the changes in technology, the challenges that they bring about, and the opportunities which they provide for demonstration and practice. However, we should clearly define our *mathematical goals*, phrase them in mathematical terms and avoid as much as possible slogans, even if we agree with their general mood. We should distinguish mathematical goals from *educational goals*. Only then may we look at issues of technology, and decide whether they help steering math education the right way, or not. To understand the effect this *process* of analyzing the role of technology has, consider the following example.

*The child will be able to derive qualitative and quantitative information form graphs such as a graph displaying the change of temperature with altitude.*

I hope everybody agrees with the statement as a basic goal of K-12 mathematical education. The terms *qualitative and quantitative information* demand further elaboration, but I shall not go into it. Now suppose we have to choose between graphic

[14] Judah Schwartz, in his essay *Intellectually stimulating and socially responsible school curricula – can technology help us get there?* writes: "By far the dominant expectation of education in most societies, at least as articulated by political leaders and by the print and electronic press, is to prepare people for the world of work."

calculators and pencil-and-paper, for a first encounter with graphs as a tool to communicate observations and measurements. Have we phrased our goal as *The child will learn to appreciate the use of graphs in natural sciences such as climatology,* we might be inclined to favor graphic calculators. They are attractive, have the fragrance of modernism, and provide vast opportunities that pencil and paper do not provide. But are they as good in conveying first principles? Can the child learn from them where to choose to draw the axes, what scale to use, and how to plot the data? Even the mere physical act, the hand-eye coordination in handling the ruler, is fundamental in my eyes to the learning process. Feeding the data into a calculator, then pressing a button, produces wonderful results, but has its pedagogical drawbacks. This does not mean I would discard graphic calculators. At a later stage they can be helpful in adding visual affects that are difficult to achieve without them – zooming in and out, changing scale, flipping the axes, to name a few. I would simply be careful in my choices, which tool to apply first in class.

While I can see the benefits of graphic calculators in middle-school in studying functional dependence, I am much less excited by the use of ordinary calculators in elementary school arithmetic. At this early stage building number sense is the teacher's number one task. I still have to hear one good argument in favor of calculators in this regard. I need no proof for how destructive they can be. Even those opposed to PPA value mental arithmetic, as means for estimation and double-check. Unfortunately we have witnessed all around us, at school and at the university, a significant decline in these skills over the last two decades, that I can only attribute to the introduction of calculators. Whoever agrees that skill building is an important component of mathematical understanding, and cannot be separated from conceptual thinking, must also confess that calculators at an early age are impeding normal mathematical development.

Those advocating early use of calculators necessarily advocate early emphasis on decimals at the expense of simple fraction arithmetic. Is it right? From the point of view of technology, simple fractions are probably obsolete. From the point of view of their pedagogical value, in understanding basic principles of arithmetic, such as ratio and proportionality, or unique factorization, and in anticipating similar structures in algebra, they are indispensable. For all these reasons I would happily ban the use of calculators in class until a solid understanding of arithmetic has been achieved, and the associated skills have been built. I am not in a position to judge whether these happen at the end of fifth, sixth or seventh grade, but the general spirit is clear to me.

Two arguments that are often heard in favor of technology at school are (a) that to oppose it is a lost battle and (b) that technological skills are so important in society, that not teaching them early would deprive certain children, especially those coming from poor families, of future opportunities. To the first argument I have nothing to say, except that if we adopt it we shouldn't be here today. As for the second, I must admit I am very sensitive to the social obligations of educators. Fortunately or unfortunately, home computers are not anymore the sign of a privileged family, much as TV is not a sign of progress, and I honestly believe that mastering Excel carries no

more mathematical value than mastering a microwave manual.

Finally, a comment on a growing trend among educators to write computer-assisted material or use sophisticated software, such as Dynamic Geometry Software, in conjunction with the standard curriculum. Some of it is very well made, enriches the learning environment, and I have no objection to computers per se. But from the little I have seen in this medium, in terms of cost-benefit analysis, the added value is not big, so I will never substitute a computer for the informal contact with a talented teacher. When it comes to political decisions, where to invest the money, my preferences are clear, at least in the country I come from.

To this one should add that computers are *not just a tool* to convey the same message more efficiently. Learning in a computerized environment affects our perception of the objects of study. Good or bad, this has to be analyzed before a new computer-dependent program is adopted.

**Tests and assessment.** Testing is a controversial issue among educators. There is a whole separate session at ICM2006 devoted to two competing international comparative tests – PISA and TIMSS. It is well known that certain educators detest testing altogether, while others build their whole curriculum around it. The more I think on it the more I become convinced that testing is a *political issue,* namely an issue that has to be decided by policy makers, based on an ideology, and taking into account factors that are only remotely related to math education. An excellent example is the controversy around US government act "No Child Left Behind" from 2002.

Testing takes various shapes. It can be comprehensive or diagnostic. You may test accumulated knowledge, or you may test the potential of a student. You may test algorithmic skills, or you may test insight and creativity. (Even though, as I said above, the former are indispensable for developing the latter, when it comes to testing, they are quite different.) A math test can be phrased in formal language or in prose. A test can be confined to one school, to one state, or to a nation. Studies show that the framework within which a problem is set affects the rate of success, and this effect changes with gender and origin. I have not mentioned more radical views which claim that western societies test only "western intelligence", and blame the relative failure of certain minorities on the dominant western frame of mind.

Testing can also serve a variety of goals. It may be purely informative, or can serve to rank, for purpose of admissions or stipends. It can test the students, but it can also test teachers success, and inform them of potential problems. Testing can be used for comparing alternative programs, or it may be needed to impose discipline on students, and on educators.

I regard all these goals as legitimate, and every kind of test welcome, *provided* one knows what kind of information to expect form it. A company recruiting civil engineers will probably test different mathematical skills than a software developer, and a matriculation exam summarizing the achievements of a student in high school need not be similar to entrance exams at a university, where a greater emphasis may be put on the student's potential and creativity.

Obviously teaching should center on the subject-matter and not only prepare for tests, but a change in curriculum often requires frequent testing to make sure the message gets across. Where there is a good tradition, and little intervention is needed, testing can be kept at a minimum. Under different circumstances tests may become a central integral part of the program.

Mathematical education will benefit from an open discussion of the issues raised here and others. It is important that mathematicians will express their views, paying respect to educators, and share their convictions with them. It is important to get to the bottom of examples, and refrain from vague statements. It is important to let changes happen, with ample time if needed, but refrain from changes that are made for the sake of reform alone. Changes must be gradual, and objectively followed. Most new ideas succeed when pushed vigorously with a small group of dedicated teachers, and with a fat budget. The problem is what happens when a case-study involving a dozen schools is over, and those ideas are adopted across the board. Do they carry enough weight to keep the momentum? Are the teachers qualified to spread the gospel?

## Concluding remarks

by *Michèle Artigue*

In this panel session, we focus on only a few of many possible controversial issues: technology, skill building, test and assessment. For each theme, as the coordinator of this panel, I articulated a short list of questions and asked the two panelists to express their positions. As could be expected, these positions are quite different, as they probably would be on the following fundamental issues: What do we want to achieve today through elementary and secondary mathematics education? What mathematics should be taught in order to achieve these goals? And how should we teach this mathematics? What are the relationships between mathematics education and the society at large?

K-12 mathematics education does not serve a unique goal. It aims at the transmission from one generation to the next one of a cultural heritage, which is one of the great achievements of humankind, and at the development of the logical reasoning competence which is so strongly attached to it. It aims at providing students with efficient means for understanding the world in which they live, and play their proper role in it. It aims at preparing and making possible the training of future mathematicians and scientists who will be in charge of the development of mathematics and scientific knowledge, and of the teachers who will have the responsibility of the transmission of this knowledge. Such ambitions can be seen as general invariants, but what is certainly not an invariant is the way we understand each of these components, at a given moment, in a given context; the way we understand the adequate balance between these, and last but not least what we consider the most appropriate strategies for achieving these ambitions. Educational systems try to adapt to this variation mainly

through curricular changes. The turbulence and controversies we regularly observe attest to the difficulty of this adaptation, and also the fact that the curricular lever chosen is not necessarily the best one.

As a mathematician who has worked in the area of mathematics education for more than 20 years now, I am struck by the simplistic way in which the complex problems that K-12 mathematics education faces today are often approached; the existing tendency to give the same value to rough affirmations and anecdotes as to well founded analysis and discussions; the persistent belief in the existence of easy and immediate solutions; the brutality of the changes imposed on educational systems, without considering their real cost, and without developing the necessary means for understanding observed success and failure. Education in the large seems a world where opposition and slogans are in front of the stage, hiding shades of meaning and dialectic visions. Slogans used by those favoring or opposing the use of technology, opposing positions on the development of concepts and of techniques are typical examples of these. Even educational research, in its attempts to reach a larger audience, does not always avoid undue simplifications and oppositions[15].

For improving the current situation, we need to overcome such a state, and will try to do so in the ICM panel associated with this contribution. But in order to solve the complex and difficult problems that K-12 mathematics education faces today in many countries, we need to do more than express well-articulated positions on controversial issues and the rationale for these. We need coherent and long-term programs, taking into account the specificities of the different contexts and the existing material and human resources. We need exchanges on our respective situations and experiences for improving these, being aware that solutions in mathematics education are always local ones in terms both of space and time, that it is nearly impossible to determine what is the exact field of validity of a given observed result, the field of extension of a given regularity. We need the collaboration of all those who are involved in mathematics education: mathematicians, mathematics educators, teachers and teachers educators, each of whom can contribute different kinds of expertise. One of the ambitions of ICMI, through its series of ICMI Studies, is to foster such exchanges among all those interested in mathematics education and to make clear what is the state of the international reflection on some selected critical issues, what has been achieved and what is needed[16].

I would like to add to these short comments that curricular choices are certainly important but that the dynamics of complex systems, such as educational systems, is not just a matter of curricular choices. The quality of teachers and of teacher education, both pre-service and also in-service, is certainly as important if not more

[15] See for instance M. Artigue, Learning Mathematics in a CAS environment: The Genesis of a Reflection About Instrumentation and the Dialectics Between Technical and Conceptual Work, *International Journal of Computers for Mathematics Learning* **7** (2002), 245–274.

[16] Themes for the most recent ICMI studies have been: The teaching and learning of mathematics at university level, the future of the teaching and learning of algebra, mathematics education in different cultural traditions – a comparative study of East Asia and the West, applications and modelling in mathematics education, the professional education and development of future teachers of mathematics.

important than curricular choices. From this point of view, the fact that mathematics educational research, which has for a long time focused on students, has in the last decade paid increasing attention to the teacher and to teacher education, is a promising evolution. Research tries today to understand the coherence underlying observed teachers' practices[17] the kind of precise mathematical knowledge the profession requires, how it can be developed, how this mathematical knowledge interacts with other forms of professional knowledge, and how these complex interactions influence teachers' practices and students' learning. Interesting results begin to be obtained, which at the same time help us understand better what can be realistic dynamics for change. The final success of the enterprise requires the collaboration of those with diverse expertise[18].

IUFM of Mathematics, Université Paris VII, France
E-mail: artigue@math.jussieu.fr

State University of New York, Buffalo, U.S.A.
E-mail: ar9@doc.ic.ac.uk

Institute of Mathematics, Hebrew University, Jerusalem, Israel
E-mail: deshalit@math.huji.ac.il

[17] A. Robert and J. Rogalski, Le système complexe et cohérent des pratiques des enseignants de mathématiques: une double approche, *La revue canadienne des sciences, des mathématiques et des technologies* **2.4** (2002), 505–528.

[18] An example of such a collaboration is given by the Mathematics and Sciences Research Institute in Berkeley which has created an education advisory board and organizes workshops involving mathematicians, mathematics educators, teachers, policy makers etc. on critical issues. The themes of the first two were the assessment of students' mathematical knowledge and the mathematics knowledge for K-8 teachers.

*Panel B*

# What are PISA and TIMSS? What do they tell us?

Lee Peng Yee (*moderator*)

Jan de Lange and William Schmidt (*panelists*)

**Abstract.** This is a panel discussion on PISA and TIMSS, two international comparative studies in educational achievement. The panelists are Jan de Lange of the Freudenthal Institute, the Netherlands, for PISA, and William Schmidt of Michigan State University, the United States, for TIMSS, with Lee Peng Yee of National Institute of Education, Singapore, as a moderator. They are to explain the nature, the aims, and the conclusions of the two studies, and to argue over their relative merits. This document contains three initial statements from the above-mentioned participants respectively.

**Mathematics Subject Classification (2000).** Primary 00A35; Secondary 00A05.

**Keywords.** PISA, TIMSS, comparative studies.

## Introduction

by *Lee Peng Yee*

One area of interest in education is comparative studies in educational achievement, in particular, in mathematics, science and reading. There are two such international studies involving mathematics, namely, PISA and TIMSS. PISA stands for the Programme for International Student Assessment. It is better known in Europe. TIMSS stands for the Trends in International Mathematics and Science Study. TIMSS was previously known as the Third International Mathematics and Science Study. Each study of PISA or TIMSS involves approximately 50 countries and thousands of students in each participating country. The studies generated volumes of publication and numerous related research projects.

The fact that some Asian countries topped the achievement list in TIMSS amazed many people and drew the attention of the industrial countries. Consequently it induced the study on these high-performing Asian countries, namely, China, Korea, Japan, and Singapore. Further a country could do well in TIMSS but not in PISA. This phenomenon is now known as PISA shock. Hence in addition people are also interested in the comparison of these two international studies. The impact of PISA and TIMSS has gone way beyond the mathematics and science community. It even

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

influences the policy makers of a country. It is timely that we have a panel discussion on the topic.

**TIMSS.** The study was commissioned by IEA, the International Association for the Evaluation of Educational Achievement. The first round of TIMSS took place in 1995 and the second round in 1999. It was the third round that made TIMSS famous world wide It collects data on educational achievement from students at the fourth and eighth grades. It also collects extensive information from students, teachers and school principals about the teaching and learning of mathematics. The test items are matched against those in the standards or syllabus. Then the data are analyzed and the reports published. The next round will take place in 2007. For details, see [1].

**PISA.** The study was initiated by the OECD countries. OECD stands for Organisation for Economic Co-operation and Development whose member countries were originally countries from Western Europe but now they are all over the globe. PISA was conducted every three years in 2000, 2003 and the next one in 2006. The tests are administered to 15-years-old students. The tests are supposed to assess how well students are prepared for their full participation in society. Similarly, the data are analyzed and the reports published. As we can see, PISA differs from TIMSS in methodology and aims. For details, see [2].

**Benchmarking.** Both PISA and TIMSS have been used by many countries for benchmarking. Roughly speaking, TIMSS is grade-based, that is, testing students of Grade 4 and Grade 8, whereas PISA is age-based, that is, testing the 15-years-old students. The 15-years-old students are those who are near the end of their compulsory education. Test items in TIMSS are more content or standards orientated, whereas those in PISA are more literacy orientated. TIMSS assesses how much students have achieved in schools. PISA assesses how well students are prepared for the outside world. Of course, this is an over-simplified view of the differences between the two studies. It does give a general idea about the two studies.

**Panelists.** They are Jan de Lange of the Freudenthal Institute, the Netherlands, speaking for PISA, and William Schmidt of Michigan State University, the United States, speaking for TIMSS. Jan de Lange is Director of the Freudenthal Institute and a full professor at University of Utrecht, the Netherlands. He was a member of the National Advisory Board of the Third International Mathematics and Science Study, and is currently Chair of the Mathematical Functional Expert Group of the OECD-PISA. William Schmidt is a professor at the College of Education, Michigan State University, and the national research coordinator and executive director of the United States National Research Center which oversees the United States' participation in the Third International Mathematics and Science Study. At the panel discussion, they are to present what PISA and TIMSS are respectively, and what they are for. Then they will discuss and possibly answer questions from the audience.

**Issues for discussion.** The issues for discussion include at least some or all of the following questions. The questions are divided into three categories. First, what are PISA and TIMSS?

- Does PISA or TIMSS really serve the purpose intended?
- Why do we need PISA when we already had TIMSS?
- What are the good points or bad points of PISA and TIMSS?

Secondly, what do they tell us?

- Is it really meaningful to use PISA or TIMSS for benchmarking?
- Some countries did well in TIMSS but not in PISA. Why?
- Both PISA and TIMSS have collected a vast amount of data. Are they useful for other researchers? What can they do with the data?

The last question above was previously raised at the International Round Table in Tokyo 2000 [3]. Thirdly, what is the future?

- The learning process of a student is a long-term affair. Perhaps the three-year cycle or four-year cycle is simply too short to measure the progressive achievement of a student. Do we need to measure so frequently?
- Will there be PISA or TIMSS 20 years from now?

This short statement serves as an introduction to the panel discussion to be held on 28 August 2006 in Madrid, Spain. Other statements from the panel speakers follow.

## References

[1] TIMSS 2003, Trends in International Mathematics and Science Study. International Mathematics reports released 14 December 2004. Website: www.timss.com.

[2] PISA 2003 technical report, OECD Programme for International Student Assessment, Website: www.pisa.org.

[3] Lee, Peng Yee, International Round Table, Proceedings of the International Congress on Mathematics Education, Tokyo 2000.

## TIMSS as a study of education: why should we care?

by *William H. Schmidt*

Comparative studies of education often seem to evoke a "so what?" or "who cares?" reaction. Studies of students' achievement in different countries may leave one wondering what practical importance such differences hold in the real world or work and commerce. Descriptions that highlight differences in common educational practices may appear intriguing and stimulate curiosity but may leave one wondering what the relevance is to what happens (or should happen) at the school down the street.

The value of such studies is almost assumed to be self-evident given, it seems, by the sort of attention the media frequently affords them. Reports of rankings along with comparisons of scores with countries x, y, and z reduce the entire endeavor to a sort of education Olympics or horserace. The value, obviously, lies in the comparisons! Who is on first? Who is doing it right?

As intriguing and entertaining as some comparisons may be – "Wow! Teachers in country x *never* assign homework!" or "Students in country z have to go to school on *Saturday*!" – these are practices that must be understood within their particular social, cultural, and educational contexts. Attempting to copy or transplant the practices of one country into another will not likely have the desired effect: alien grafts rarely take without extensive preparation and effort.

**Value of international comparative studies.** The real value in international studies lies not in the comparisons themselves, but in the insights we may gain into our own common practices. International comparisons hold up and frame what's familiar against a background of a considerable range of alternatives. This can lead to a thoughtful reconsideration of our rationale for doing things the way we do – or even initiate a thoughtful evaluation of something never before considered.

Many people, for example, are familiar with schools. They know what schools are and what happens in the classrooms inside the schools. Schools are schools; whether they are urban Paris or rural Montana. We began our involvement with international education research in the days leading up to TIMSS with a similar assumption about the nature of schools in various countries. We discovered that school has commonalities everywhere. What is common wherever schools are found are students, teachers, and textbooks. How these commonalities interact and work within a larger education system, however, can vary considerably. We discovered, for example, that in Norway primary teachers typically stay with the same group of students for the first five or six years of students' formal school experience. We also learned that in Switzerland, 'schools' only exist in large cities. The majority of students and teachers meet together in rooms located in buildings that are not necessarily dedicated to housing educational activities. Furthermore, school administrators and other support personnel are only found in such dedicated facilities which generally house the upper secondary grades or are located in the cities.

Clearly there are a number of ways to conduct school. The examples mentioned here were not selected to suggest that all countries change either the nature of their school buildings nor the length of time primary teachers typically work with the same group of students. Some consideration of these issues may be fruitful, but the point to be made here is that these schooling practices represent options – choices that have been made about how school gets done. The more we can see the way we do things as choices, the better position we are in to consider and construct profitable change or reform.

**What we can learn from TIMSS.** In the Third International Mathematics and Science Study (TIMSS) the focus was not on the structural aspects of school such as the previously mentioned examples, although these were a part of the study. The focus, rather, was on the substance of education, the school curriculum, the content that's at the center of what teachers and students do in schools.

Previous international studies led us to suspect that the achieved curriculum, what students demonstrate that they know, varies from one country to another. TIMSS assessed this aspect of curriculum in the context of an extensive investigation of the intended curriculum, what systems intend their students to learn, along with the implemented curriculum, what is taught in the classroom. Measurements of these curriculum aspects led to one unmistakable conclusion: the mathematics taught and studied in the schools of one country can differ substantially from what exists in the schools of another. In short, there are many ways to do mathematics education.

More specifically, this curriculum measurement in TIMSS led to some thought provoking insights into the U.S. mathematics curriculum. For example, the U.S. intends teachers and students to study two to three times the number of topics in the first through eighth grade as is typical in other countries. Consistent with this breadth, U.S. textbooks are truly first in the world in their size, weight, and scope. Not too surprisingly, given these intentions and resources, the U.S. teachers tend to spend some time on every intended topic typically without emphasizing any small number of topics as is common in other countries. All of this contributes to the "mile wide, inch deep" nature of the U.S. curriculum.

These insights were possible because TIMSS was designed from the start to examine the relationship among the various aspects of the curriculum: the intended, the implemented, and the attained. These insights have also led to several efforts to thoughtfully revise the U.S. mathematics curriculum.

So, what is the value of international study? Certainly not to obtain bragging rights for the top spot on some list nor even to identify specific practices that we may want to copy. The real value stems from obtaining a fresh perspective on the array of choices embedded in our own approach to education. Thoughtful and principled insights stimulated by examples from other systems can lead to powerful revision in our quest to provide a challenging and equitable education for all students.

## PISA: promises, problems and possibilities

by *Jan de Lange*

**PISA versus TIMSS.** According to the OECD:

> The OECD's Programme for International Student Assessment (PISA) is a collaborative effort among the member countries of the OECD to measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge societies. The assessment is forward looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum. This orientation reflects a change in the goals and objectives of curricula themselves, which are increasingly concerned with what students can do with what they learn at school, and not merely whether they have learned it. The term 'literacy' is used to encapsulate this broader conception of knowledge and skills.

The first PISA survey was carried out in 2000 in 32 countries, including 28 OECD member countries. Another 13 countries completed PISA 2000 in 2002, and from PISA 2003 onwards more than 45 countries will participate 'representing more than one third of the world population'. PISA 2000 surveyed reading literacy, mathematical literacy, and scientific literacy, with the primary focus on reading. In 2003 the main focus was on mathematical literacy (published in 2004), and in 2006 scientific literacy will be highlighted.

It will be clear that TIMSS and PISA have a lot of similarities resulting in improper identification of the two series of studies in the media, which is undesirable and confusing. But the descriptions of the organizations that are responsible, show that they both claim similar relevance for the studies. Even for the expert it will be difficult to relate the following either to TIMSS or to PISA: 'Countries participating in this study will have information at regular intervals about how well their students read and what they know and can do in mathematics and science.' Both studies do this and do it, methodologically speaking, in a very similar way (based on Item Response Theory, IRT). Even the reporting tables in the respective reports look very similar.

If there is a problem that both studies share, it is the design of the measuring instrument in relation to the validity of the outcomes. Traditionally, validity concerns associated with tests have centered about test content, meaning how the subject domain has been sampled. Typically evidence is collected through expert appraisal of alignment between the content of the assessment tasks and the curriculum standards (in case of TIMSS) and 'subject matter' assessment framework (PISA). Nowadays, empirical data are often used before an item is included in a test.

Traditionally validation emphasized consistency with other measures, as well as the search for indirect indicators that can show this consistency statistically. More

recently is the recognition that these data should be supplemented with evidence of the cognitive or substantive aspect of validity. Or as *Knowing What Student Knows* (2001) summarized: 'The trustworthiness of the interpretation of test scores should rest in part on empirical evidence that the assessment tasks actually tap the intended cognitive process.'

One method to do this is a protocol analysis in which students are asked to think aloud as they solve problems; another is an analysis of reasons in which students are asked to provide rationales for their responses; and a third method is an analysis of errors in which one draws inferences about processes from incorrect procedures, concepts, or representations of problems. Although some of these methods are applied only after the test is administered, there is a trend that large-scale assessments like TIMSS and PISA use these methods as well. The use of cognitive laboratories to gauge whether students respond to the items in ways the developers intended has become a new instrument in the developmental process. The use of double-digit coding is another sign of interest in the process of problem solving instead of just judging whether an answer is incorrect or correct. A 'correct' or 'partly correct' score given not only to each work of the student, but also to which strategy was used or where in the process the students 'lost track'.

**Validity.** The validity of the test instrument remains a complex issue. It goes without saying that there is an inherent tension between the traditional choice of item formats, usually with very restricted time (1–2 minutes per item), and the rather ambitious definitions of what the instrument is intended to measure. But not only the concern about 'errors' plays an important role in relying so much on multiple-choice, it is also an economic issue: Many countries participating in these large cooperative studies are unwilling or unable to fund much more expensive multiple marker studies, even if such studies have demonstrated their efficacy.

PISA 2003 also had a problem solving component. Many of the items would fit the mathematics Framework, and given the fact that the instrument for problem solving (PS) had much more open 'constructive' items, a study relating the math items and the PS items could be very helpful in advancing the discussion on item instruments and their restrictions in large-scale international studies. According to the PISA report on problem solving: 'The items for problem solving give a first glimpse of what students can do when asked to use their total accumulated knowledge and skills to solve problems in authentic situations that are not associated with a single part of the school curriculum.'

One can easily argue that this is always the case in a curriculum: For mathematical literacy, mathematics as taught at school will not suffice. Students need to read, need to interpret tables and graphs (seen by many as belonging to reading literacy), and, indeed, need problem-solving strategies. But seen from the perspective of promising developments on item formats and item quality, the problem-solving component of PISA is interesting, at least. And if TIMSS implements their intent to 'place more emphasis on questions and tasks that offer better insights into students' analytical,

problem-solving, and inquiry skills and capabilities,' innovation in large-scale assessments could materialize.

**PISA versus TIMSS.** The main differences between TIMSS and PISA seem to be the following:

- curricular emphasis for TIMSS versus functional aspect (literacy) for PISA;
- grade-specific structure of TIMSS versus age-specific structure of PISA.

TIMSS uses the curriculum as the major organizational aspect. The TIMSS curriculum model has three aspects: the intended curriculum, the implemented curriculum, and the achieved curriculum. These represent, respectively, the mathematics and science intended for students to learn, and how the education system should be organized to facilitate this learning: what is actually taught in the classrooms, who teaches it, and how is it taught; and finally, what it is that students have learned, and what they think about those subjects.

International curricular diversity was a serious point of concern to the TIMSS study. The goal was to develop an international test that would be equally fair to all participating countries. Therefore subject-matter specialists from all countries were consulted and asked to contribute to the process of test development. Most countries participating in TIMSS had an intended mathematics curriculum that matched with more than 90% of the items. The outliers were the United States and Hungary with 100% matching, and the Netherlands, with 71% matching.

Insiders have discussed the procedure and its validity of this equally unfair analysis. The question not satisfactorily answered is how the mathematics education communities in the different countries were involved, and how representative they were. But if these numbers are accepted, in this context it is worth looking at the minimal matching result of the Netherlands.

It was expected that students of other countries would outperform Dutch students. However, contrary to expectations, in 1995 Dutch grade 8 students performed well on the TIMSS test. Their score was significantly above the international average, just below the four Asian top-scoring countries. After some additional research it was concluded that somehow the Dutch students were knowledgeable about the 29% of test items that were remote from their intended curriculum. In the end it was concluded that the students had the abilities for transfer of their knowledge and skills to items that did not match with their intended curriculum. It can be very appropriate to test students on material they have not been taught, if the test is used to find out whether the schools are doing their job.

PISA takes this point even further: It is based on a dynamic model of lifelong learning in which new knowledge and skills necessary for successful adaptation to a changing world are continuously acquired throughout life. It focuses on young people's ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum.

The two different approaches can both be critiqued: What does it mean that the Netherlands scored so high with the minimal relation with its curriculum? What does

it mean if PISA will not constrain itself to any national curricula? It is clearly not true that international studies of student achievement may be unintentionally measuring little more than the degree of alignment between the test instrument and the curriculum. What it does measure is still a question open to interpretation.

Another indication that shows how difficult it is to make statements that go beyond well-intended opinions can be found in the observation of Westbury in 1992, in relation to SIMS, when he observed that the lower achievement of the United States is the result of curricula that are not as well matched to the SIMS test as are the curricula of Japan. But in TIMSS the match was 100% (see earlier), and still the United States did not perform very well.

**Impact.** The Germans produced a national PISA 2000 report of 550 pages, the international OECD report was 330 pages, and the Dutch report a mere 65 pages. Most countries had something around 150 pages. It is not the statistics that are interesting here, but the message from the report and what has been selected to be included. Even a superficial analysis, which was carried out for this article with the reports mentioned and the one from the United States, makes significant differences visible. There is a common myth that numbers do not lie. It is now widely accepted that data can be gathered, processed, mathematized, and interpreted in a variety of ways. So a key issue is the question of who influences this process, for what reasons, and through what means. The studies just mentioned underscore this concern apart from the fact that even numbers can lie.

Back to the very *gründliches* German report. Not only did the German PISA Konsortium do an excellent and thoughtful job, it also made recommendations for immediate improvement, including ones that directly affect the content. The changes should include:

- more integration of inner- and outer-mathematical 'networks';
- fewer calculations;
- more thinking activities and student mental 'constructions';
- more reflection;
- more flexible use of schoolbooks.

These goals can be reached when the recommendations that were formulated after TIMSS are implemented:

- development of a different math-problems culture: more open-ended, more 'real-world';
- a new teaching-and-learning culture, with a more exiting cognitive school environment;
- more and different professionalization of teachers, emphasizing teamwork.

PISA adds to these recommendations a 'very different conceptualization' of mathematical concepts and emphasis of modeling and mathematization, situated in contexts. And, argued the report, the Germans have definitely not reached the optimum in using different representations as a tool to build better conceptual understanding.

Mathematics education is in a state of transition, in part because of the fact that both TIMSS and PISA were taken seriously. Surprisingly the shock and catastrophe that struck Germany as some kind of natural disaster, if one had only the popular media as a resource, has resulted in a government-supported nationwide action-plan with a very strong content part that will result in a different mathematics education culture at schools. Of course, the success of these changes will be measured by PISA 2003, 2006, 2009, and so on. At least in part.

**The future of PISA.** It is very hard to predict the future of PISA. Of course it is a very successful project if one looks at the number of countries participating: 58 in 2006 and growing. And there are many opportunities to make PISA more successful from the content point of view. If PISA is able to include longer and more complex items, as it did with its Problem Solving study in 2003, if technology gets a proper place (as is intended), if group-work can be included in some way PISA would make itself much more rewarding for policy makers and practitioners alike.

PISA will also start a study for the 9-year olds, in the near future. In short the OECD definitely has the intention to continue PISA for the next decade at least. And if the instrument keeps improving, it seems worth the effort – although OECD has to be more clear about the fact that PISA measures mathematical literacy, and not curricular mathematics – and how to deal with this principle in the future.

PISA will have to address the problem of the Horse-race – a very undesirable aspect that draws a lot of criticism – and rightfully so. Another format of the international report with portraying country by country would not only be more informative, but also would give a more valid picture: one number cannot represent the quality of an educational system.

Validity issues have to be addressed, even if PISA is using state-of-the-art methodology. Not only the methodology should be of the highest quality, but also the content – and improvement should be on the agenda continuously.

And of course: communication between all parties should improve: Math educators and research mathematicians feel as being watchers of a game they hardly feel any ownership for. This is undesirable: PISA should not address just policy makers if it really wants to make a difference: the data of PISA are in the public domain and any country can analyse these data for its own purpose. This opportunity should not be lost. The meaning of PISA can be co-defined by its users.

National Institute of Education of Singapore, Singapore 637616, Singapore
E-mail: pylee@nie.edu.sg

Freudenthal Institute, University of Utrecht, 3561 GE Utrecht, The Netherlands
E-mail: J.deLang@fi.uu.ni

Center for the Study of Curriculum, Michigan State University, East Lansing, MI 48824-1034, U.S.A.
E-mail: bschmidt@msu.edu

*Panel C*

# The role of mathematicians in K-12 mathematics education

Fr. Ben Nebres (*moderator*)

Shiu-Yuen Cheng, Konrad Osterwalder, and Hung-Hsi Wu (*panelists*)

**Abstract.** The need for mathematics educators, schoolteachers and mathematicians to work together to improve K-12 mathematics education continues to be a great concern throughout the world. The main paper for this panel discussion proposes a paradigm or perspective within which to organize this working together of the different groups. This is to see mathematics education as mathematical engineering. From this perspective, the challenge of the mathematics educator and the schoolteacher is to customize mathematics to students' needs. The role, in turn, of the university mathematician is to customize mathematics courses for teachers so that they in turn may be able to customize the mathematics for the different needs of their students. This paradigm is then discussed in different contexts, in the United States, Hong Kong, Switzerland, and the Philippines. The paradigm is seen to be fruitful in these different contexts.

**Mathematics Subject Classification (2000).** Primary 97D20; Secondary 97B99.

**Keywords.** Mathematical Engineering, curriculum, assessment, professional development, student creativity, gymnasium mathematics.

## Introduction and overview

by *Ben Nebres, S. J.*

The theme of this panel discussion is "How mathematicians contribute to K-12 mathematics education." Three distinguished mathematicians, coming from different contexts and different mathematics education traditions, provide challenging and helpful insights into this theme. Because the contexts in which they write are quite different (United States, China, Switzerland), it was decided to present the papers separately. This introductory note is meant to highlight the main proposal and perspective coming from Prof. Wu's paper and relate the contributions from the other two panelists to it. I also add a few comments from the context of a developing country, the Philippines.

First, a brief note on the diversity of contexts. In terms of mathematics achievement based on international comparative studies such as TIMSS, the United States ranks in the middle, while Hong Kong and Switzerland rank towards the top. The Philippines ranks towards the bottom. In terms of educational systems, the U.S. is quite decentralized with great diversity in terms of curriculum, textbooks, teacher

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

training, while Hong Kong schools would have greater commonality in terms of curriculum, textbooks, assessment. Some would say that it may be better to compare not the performance of all U.S. schools, but to take account of the diversity of systems and compare blocks of schools (by states or groups of school districts). In terms of resources, the U.S., Switzerland and Hong Kong have first world resources, while Philippine schools operate in the context of great scarcity: classes of 80 students in rooms built for 40, one textbook shared by 5 or 6 pupils and so forth. One can even look at the differences in mathematics education between Hong Kong and the U.S. and Switzerland and the Philippines from the point of view of mathematics education cultures. This is discussed in the recently published "Mathematics Education in Different Cultural Traditions: A Comparative Study of East-Asia and the West", edited by Frederick Leung, Klaus Dieter Graf and Frances Lopez-Real, Volume 9 in the New ICMI Study Series. My own role as a mathematician in helping improve mathematics education in the Philippine context of poverty of resources is described in a chapter entitled "Philippine Perspective on the ICMI Comparative Study" in this volume of the ICMI Study Series.

Despite this diversity of contexts, there is agreement in all the three papers (and in the Philippine experience as well) on the importance of the role of mathematicians in K-12 mathematics education and on a particular paradigm or perspective (mathematics education as mathematical engineering) on how mathematicians can effectively fulfill this role.

In the main paper for this panel presentation and discussion, Prof. Hung-Hsi Wu of the University of California Berkeley, proposes a re-conceptualization of mathematics education as mathematical engineering: "Thus chemical engineering is the science of customizing chemistry to solve human problems… . I will put forth the contention that mathematics education is mathematical engineering, in the sense that it is the application of basic mathematical principles to meet the needs of teachers and students." This is somewhat different from the suggestion of Hyman Bass to look at mathematics education as a branch of applied mathematics. In engineering, what is important is the customization of scientific principles to address human needs. Similarly, in mathematics education as mathematical engineering, what is crucial is the customization of mathematical principles to address the needs of teachers and pupils.

From this viewpoint, the challenge is to work out the role of mathematicians in mathematics education analogous to that of physicists in engineering. Just as the roles of physicists and engineers in engineering are deeply intertwined, so should the roles of mathematicians and mathematics educators be in mathematics education. Right now the two worlds are separate and do not communicate well with each other. Prof. Wu writes: "… if mathematicians want to participate in serious educational work in K-12, … the most important thing is the awareness that K-12 mathematics education is not a subset of mathematics, and that there is quite a bit to learn about the process of customization that distinguishes K-12 mathematics education from mathematics."

In my communication with Prof. Wu, we agreed that it is important that the term "mathematics educator" include both the university mathematics educator as

researcher and the school mathematics teacher as practitioner. While the university mathematics education researcher is an expert on teaching and learning theories, the mathematics master teacher is most knowledgeable about actual teacher, student, and classroom contexts. Success in improving mathematics education will require good communication and working together among mathematicians, university mathematics educators and school mathematics teachers.

To properly customize mathematics in different student contexts, a mathematics teacher needs: solid mathematical knowledge, clear perception of the setting defined by students' knowledge, and flexibility of mind to customize this mathematics knowledge for use in this particular setting. In this model of mathematics education as mathematical engineering, the role of the mathematician is to provide the solid mathematical knowledge. This should be done in such a way that the teacher is provided with different ways of understanding and approaching a mathematics concept so that he can have a repertoire to draw from in customizing the mathematics for different student contexts. Prof. Wu gives examples such as in the teaching of fractions or in providing intervention for students at-risk.

The paper of Prof. Shiu-Yuen Cheng of the Hong Kong University of Science and Technology picks up from the "mathematics education as mathematical engineering" framework of Prof. Wu and sets it in the context of mathematics education in Hong Kong. He notes "the main factors for providing an effective mathematics education as curriculum design, teacher competence and assessment methods." He says that most important is teacher competence and it is to this factor that mathematicians can contribute the most. They can contribute in the university curriculum for mathematics teacher programs and in in-service workshops for mathematics teachers. Together with Prof. Wu (and from my experience as well) he points out that the university curriculum for mathematics teachers, which is usually a combination of courses for mathematics majors and education courses, "do not serve the purpose of providing the necessary understanding to be a competent mathematics teacher." (Reasons for this are very well argued in Liping Ma, "Knowing and Teaching Elementary Mathematics.")

Prof. Cheng points out that the Hong Kong mathematics education context is one which shows great success as shown by the excellent performance of Hong Kong students in international comparative studies. "In Hong Kong, Johnny can add! In fact, Johnny can do fractions and decimals quite well." There is, however, a downside to this achievement. Prof. Cheng is concerned that this is at great cost, particularly in "suffocating students' creativity and motivation for learning." He stresses the important role of mathematicians in communicating effectively to the public and to decision-makers these important concerns for mathematics education. (This balance between effective mastery of fundamentals and the need to foster creativity has been an important recent concern in East Asia and was the theme of the ICMI-East Asian Regional Conference on Mathematics Education in Shanghai in August 2005.)

Prof. K. Osterwalder of the ETH Zurich writes in the context of the upper years of the Swiss gymnasium (years 9–12) and the role of mathematicians in universities

such as ETH Zurich in preparing mathematics teachers for these upper years. He points out that in Switzerland, students do quite well in international comparative studies in mathematics. Teachers are well trained. They are required to get a masters degree at a level where they could equally opt to go into industry as mathematicians. On the role of mathematicians in K-12 mathematics education, he agrees that the main contribution of research mathematicians is in the education of mathematics teachers. He focuses in a special way on the "Specialized Mathematics Courses with an Educational Focus" taken by mathematics teachers in the university. He provides various examples of course material, from linear equations and linear algebra to noting recent research breakthroughs accessible to gymnasium students, where these course materials "narrow the gap between Gymnasium mathematics and University mathematics" in the spirit of Felix Klein's "Elementary Mathematics from a Higher Viewpoint."

The paradigm or perspective of "mathematics education as mathematical engineering" proposed by Prof. Wu is thus seen to be quite fruitful in these different contexts. They all point to the central role of the mathematics teacher, whose challenge is to customize the mathematics to the students' needs. The important role of the university mathematician is then to customize mathematics courses for these teachers in such a way that they may in turn be able to customize the mathematics needed by students in different contexts and with different needs.

## How mathematicians can contribute to K-12 mathematics education

by *Hung-Hsi Wu*

> *"To overcome the isolation of education research, more effective links must be created between educational faculties and the faculties of universities. This could allow scholars of education better acquaintance with new developments in and across the disciplines and other professional fields of the university, while also encouraging discipline-based scholars with interests in education to collaborate in the study of education."*
>
> Lagemann [14], p. 241.

I would like to make a general disclaimer at the outset. I think I should only talk about things I know firsthand, so I will limit my comments to the K-12 mathematics education in the U.S. rather than take a more global view. Such a restriction is not necessarily fatal since a friend of mine observed that what takes place in the U.S. tends also to take place elsewhere a few years later. For example, in France there is now a Math War that resembles the American Math Wars of the nineties (Education Week [7]). We live in a global village after all.

Let me begin with a fairy tale. Two villages are separated by a hill, and it was decided that for ease of contact, they would drill a tunnel. Each village was entrusted

with the drilling of its own half of the tunnel, but after both had done their work, it was discovered that the two halves didn't meet in the middle of the hill. Even though a connecting tunnel between the two lengths already built could be done at relatively small expense, the two villages, each in defense of its honor, prefer to continue the quarrel to this day.

This fairy tale is too close to reality for comfort when the two villages are replaced by the education and mathematics communities, with the former emphasizing the overriding importance of pedagogy and the latter, mathematical content.[1] Mathematics education rests on the twin pillars of mathematics and pedagogy, but the ongoing saga in mathematics education is mostly a series of episodes pitting one against the other. There is probably no better proof of the disunity between these communities than the very title of this article. Indeed, if someone were to write about "How chemists can contribute to chemical engineering", that person would be considered a crank for wasting ink on a non-issue. Chemical engineering is a well-defined discipline, and chemical engineers are perfectly capable of doing what they are entrusted to do. They know the chemistry they need for their work, and if there is any doubt, they would freely consult with their colleagues in chemistry in the spirit of cooperation and collegiality. Therefore, the fact that we are going to discuss "How mathematicians can contribute to K-12 mathematics education" in the setting of the International Congress speaks volumes about both mathematics education and mathematicians.

In matters of education it is of course natural for the power structure to hold the reins, just as in matters of engineering they are held by engineers. But while the chemical engineers are glad to have chemists down the hall, and glad to learn what they can use in their work, the corresponding relationship has not been the case for mathematics educators. Since education research is thriving and research funding is ample, it is not surprising that educators want to protect their intellectual independence in the university environment. Rumblings about how mathematically unqualified teachers or deficient curricula are undercutting mathematics learning do surface from time to time, but we have not witnessed the expected aggressive action agitating for collaboration with mathematicians. Other troubling issues related to mathematics content, such as the presence of incorrect assessment items in standardized tests, likewise fail to arouse genuine concern in the mathematics education community. To an outsider, the protection of the "education" enclave seems to matter more to university educators than collaboration with the research mathematics community that could strengthen K-12 mathematics education. By contrast, if the department of chemical engineering consistently produces engineers with a defective knowledge of chemistry, or if accidents occur in its laboratories with regular frequency, would the chemical engineering faculty not immediately spring to action? This question prompts the thought that maybe we no longer know what mathematics education is

[1] In writing about sociological phenomena, especially education, it is understood that all statements are statistical in nature unless stated to the contrary, and that exceptions are part and parcel to each statement. In fact, there are striking (though isolated) exceptions in the present context. The reader is asked to be aware of this caveat for the rest of this article.

about and it is time for us to take a second look.

One meaning of the word "engineering" is the art or science of customizing scientific theory to meet human needs. Thus chemical engineering is the science of customizing chemistry to solve human problems, or electrical engineering is the science of customizing electromagnetic theory to design all the nice gadgets that we have come to consider indispensable. I will put forth the contention that mathematics education is mathematical engineering, in the sense that it is the customization of basic mathematical principles to meet the needs of teachers and students.[2] I will try to convince you that this is a good model for the understanding of mathematics education before proceeding to a discussion of how mathematicians can contribute to K-12 mathematics education. The far-from-surprising conclusion is that, unless mathematicians and educators can work as equal partners, K-12 mathematics education cannot improve.

Regarding the nature of mathematics education, Bass made a similar suggestion in [5] that it should be considered a branch of applied mathematics.[3] What I would like to emphasize is the aspect of engineering that customizes scientific principles to the needs of humanity in contrast with the scientific-application aspect of applied mathematics. Thus, when H. Hertz demonstrated the possibility of broadcasting and receiving electromagnetic waves, he made a breakthrough in science by making a scientific application of Maxwell's theory. But when G. Marconi makes use of Hertz's discovery to create a radio, Marconi was making a fundamental contribution in electrical engineering, because he had taken the extra step of harnessing an abstract phenomenon to fill a human need.[4] In this sense what separates mathematics education as mathematical engineering from mathematics education as applied mathematics is the crucial step of customizing the mathematics, rather than simply applying it in a straightforward manner to the specific needs of the classroom. There is no better illustration of this idea of customization than the teaching of fractions in upper elementary and middle schools, as I now explain.

Students' failure to learn fractions is well-known. School texts usually present a fraction as parts of a whole, i.e., pieces of a pizza, and this is the most basic conception of a fraction for most elementary students. However, when fractions are applied to

[2]After the completion of this article, Skip Fennell brought to my attention the article "Access and Opportunities to Learn Are Not Accidents: Engineering Mathematical Progress in Your School" by William F. Tate, which is available at: http://www.serve.org/_downloads/publications/AccessAndOpportunities.pdf. Tate is concerned with equity and uses "engineering" as a metaphor to emphasize the potential for designing different educational policies and pedagogical activities to promote learning, but without addressing the mathematics. On the other hand, the present article explains why mathematics education is the engineering of mathematics.

[3]Hy Bass lectured on this idea in December of 1996 at MSRI, but [5] seems to be a convenient reference. After the completion of this article, Zalman Usiskin informed me that in the Proceedings of the U.S.-Japan workshop on the mathematics education of teachers in 2000 that followed ICME-9 in Japan, he had written that " 'Teachers' mathematics' is a field of applied mathematics that deserves its own place in the curriculum." Along this line, let it be mentioned that the paper of Ferrini-Mundy and Findell [8] made the same assertion and, like Bass, it does not touch on the engineering aspect of mathematics education. The need for mathematicians and educators to work on equal footing in mathematics education is likewise not mentioned by these educators.

[4]The invention was actually due to N. Tesla, but like many things in life, popular preception displaces the truth. I am indebted to S. Simic for pointing this out to me.

everyday situations, then it is clear that there is more to fractions than parts-of-a-whole, e.g., if there are 15 boys and 18 girls in a classroom, then the ratio of boys to girls is the fraction, which has nothing to do with cutting up a pizza into 18 equal parts and taking 15. In the primary grades, it is not a serious problem if students' knowledge of fractions is imprecise and informal, so that a fraction can be simultaneously parts-of-a-whole, a ratio, a division, and an operator[5], and a number. Children at that age are probably not given to doubts about the improbability of an object having so many wondrous attributes. At some stage of their mathematical development, however, they will have to make sense of these different "personalities" of a fraction. It is this transition from intuitive knowledge to a more formal and abstract kind of mathematical knowledge that causes the most learning problems. This transition usually takes place in grades 5–7.

There is by now copious mathematics education research[6] on how to facilitate children's learning of the fraction concept at this critical juncture in order to optimize their ability to use fractions efficiently. At present, what most children get from their classroom instruction on fractions is a fragmented picture of a fraction with all these different "personalities" lurking around and coming forward seemingly randomly. What a large part of this research does is to address this fragmentation by emphasizing the cognitive connections between these "personalities". It does so by helping children construct their intuitive knowledge of the different "personalities" of a fraction through the use of problems, hands-on activities, and contextual presentations.

This is a good first step, and yet, if we think through students' mathematical needs beyond grade 7, then we may come to the conclusion that establishing cognitive connections does not go far enough. What students need is an unambiguous definition of a fraction which tells them what a fraction really is. They also need to be exposed to direct, mathematical, connections between this definition and the other "personalities" of a fraction. They have to learn that mathematics is simple and understandable, in the sense that if they can hold onto one clear meaning of a fraction and can reason for themselves, then they can learn all about fractions without ever being surprised by any of these other "personalities".

From a mathematician's perspective, this scenario of having to develop a concept with multiple interpretations is all too familiar. In college courses, one approaches rational numbers (both positive and negative fractions) either abstractly as the prime field of characteristic zero, or as the field of quotients of the integers. The problem is that neither is suitable for use with fifth graders. This fact is recognized by mathematics education researchers, as is the fact that from such a precise and abstract definition of rational numbers, one can prove all the assorted "personalities" of rational numbers. If I have read the research literature correctly, these researchers despair of ever being able to offer proofs once they are forced to operate without an abstract definition, and

[5]For example, the fraction can be regarded as a function (operator) which associates to each quantity three-quarters of the same quantity.

[6]Here as elsewhere, I will not supply explicit references because I do not wish to appear to be targeting specific persons or works in my criticism. I will be making generic comments about several general areas.

that is why they opt for establishing cognitive, rather than mathematical connections among the "personalities" of rational numbers. The needs of the classroom would seem to be in conflict with the mathematics. At this point, engineering enters.

It turns out that, by changing the mathematical landscape entirely and leaving quotient fields and ordered pairs behind, it is possible to teach fractions as mathematics in elementary school, by finding an alternate mathematical route around these abstractions that would be suitable for consumption by children in grades 5-7. Without going into details, suffice it to say that at least the mathematical difficulties can be overcome, for example, by identifying fractions with certain points on the number line (for this systematic development, see, e.g., Jensen [11], or Wu [25]). What is of interest in this context is that this approach to fractions is specific to the needs of elementary school and is not likely to be taught, ever, in any other situation. In addition, the working out of the basic properties of fractions from this viewpoint is not quite straightforward, and it definitely requires the expertise of a research mathematician. As to the further pedagogical implementation to render such an approach usable in grades 5–7, the input of teachers and educators would be absolutely indispensable.[7] We therefore get to witness how mathematicians and educators are both needed to turn a piece of abstract mathematics into usable lessons in the school classroom. This is customization of abstract theory for a specific human need, and this is engineering at work.

Through this one example of fractions, we get a glimpse of how the principles of mathematical engineering govern the design of a curriculum. Less obvious but of equal importance is the fact that even mathematics education research cannot be disconnected from the same principles. If, for example, a strong mathematical presence had been integral to the research on fractions and rational numbers, it would be very surprising that the research direction would have developed in the direction it did. Compare the quote by Lagemann at the beginning of this article as well as Lagemann [14].

An entirely analogous discussion of customization can be given to any aspect of mathematics education, but we single out the following for further illustrations:

(a) The design of an "Intervention Program" for at-risk students. Up to this point, the methods devised to help these students are largely a matter of teaching a watered-down version of each topic at reduced pace; this is poor engineering from both the theoretical and the practical point of view. In Milgram-Wu [18], a radically different mathematical engineering design is proposed to deal with this problem.

(b) The teaching of beginning algebra in middle school. The way symbols are usually handled in such courses, which necessitates prolix discussions in the research literature of the subtlety of the equal sign, and the way variable is introduced as the central concept in school algebra are clear indications that the algebra we teach students at present has not yet been properly customized

[7] Some teachers who have worked with me are trying out this approach with their students in San Francisco.

for the needs of school students. See the Preface and Sections 1 and 2 of Wu [30], and also Wu [31], for a more detailed account of both the problems and their proposed solutions.

(c) The writing of mathematics standards at the national or state level. This is an example of what might be called "practical optimization problems", which customize the mathematics to meet diverse, and at times conflicting, needs of different clientele. Cf. Klein [13].

The concept of mathematics education as mathematical engineering also sheds some light on Lee Shulman's concept of pedagogical content knowledge ([20]). There has been a good deal of interest in precisely describing the kind of knowledge a teacher should possess in order to be effective in teaching. In the field of mathematics, at least, this goal has proven to be elusive thus far (but cf. Hill-Rowan-Ball [9]), but Shulman's intuitive and appealing formulation of this concept crystallizes the diverse ideas concerning an essential component of good teaching. From the point of view of mathematical engineering, one of the primary responsibilities of a teacher is to customize her mathematical knowledge in accordance with the needs of each situation for students' consumption. This particular engineering knowledge is the essence of pedagogical content knowledge. Although this approach to pedagogical content knowledge does not add anything new to its conception, it does provide a framework to understand this knowledge within mathematics, one that is different from what one normally encounters in educational discussions. It makes explicit at least three components to effective teaching: a solid mathematical knowledge, a clear perception of the setting defined by students' knowledge, and the flexibility of mind to customize this mathematical knowledge for use in this particular setting without sacrificing mathematical integrity.

The idea of customizing mathematics "without sacrificing mathematical integrity" is central to mathematical engineering. In engineering, it is obvious that, in trying to customize scientific principles to meet the needs of humanity, we cannot contradict nature regardless of how great the human needs may be. In other words, one respects the integrity of science and does not attempt anything so foolish as the construction of anti-gravity or perpetual-motion machines. Likewise, as mathematical engineering, mathematics education accepts the centrality of mathematics as a given. Again using the example of teaching fractions, a mathematics educator would know that no matter how one tries to teach fractions, it must be done in a way that respects the abstract meaning of a fraction even if the latter is never used explicitly. If, for instance, an educator catches himself saying that children must adopt new rules for fractions that often conflict with well-established ideas about whole numbers, then he knows he is teaching fractions the wrong way because, no matter what efforts one puts into making fractions intuitive to children, one cannot do violence to the immutable fact that the rational numbers contain the integers as a sub-ring. The need to teach the arithmetic of fractions as a natural extension of the arithmetic of whole numbers has gone unnoticed for far too long, with the result that too many of our students begin to

harbor the notion that, after the whole numbers, the arithmetic of fractions is a new beginning. Such bad mathematical engineering in curricular designs is unfortunately a common occurrence.

The only way to minimize such engineering errors is to have both mathematicians and educators closely oversee each curricular design. In fact, if we believe in the concept of mathematics education as mathematical engineering, then the two communities must work together in all phases of mathematics education: Any education project in mathematics must begin with a sound conception of the mathematics involved, and there has to be a clear understanding of what the educational goal is before one can talk about customization. In this process, there is little that is purely mathematical or purely educational; almost every step is a mixture of both. Mathematics and education are completely intertwined in mathematical engineering. Mathematicians cannot contribute to K-12 mathematics education if they are treated as outsiders.[8] They have to work alongside the educators on equal footing in the planning, implementation, and evaluation of each project. But this is far from the reality at present.

For at least three decades now, the mathematics and K-12 education communities in the U.S. have not been on speaking terms in the figurative sense. (Cf. Washington Post [21].) The harm this communication gap has brought to K-12 mathematics education can be partially itemized, but before doing that, let me point out three general consequences of a philosophical nature. The first one is that the isolation of the education community from mathematicians causes educational discussions to over-focus on the purely education aspect of mathematics education while seemingly always leaving the mathematics untouched. The result is the emergence of a subtle mathematics avoidance syndrome in the education community, and this syndrome will be seen to weave in and out of the following discussion of the specific harmful effects of this communication gap. Given the central position of mathematics in mathematical engineering, it would be noncontroversial to say that this syndrome should vanish from all discussions in mathematics education as soon as possible.

The fact that many mathematicians teach mathematics and design mathematics courses throughout their careers seems to escape the attention of many educators. Here is a huge reservoir of knowledge and experience in mathematical engineering on tap. The chasm between the two communities in effect denies educators access to this human resource at a time when educators need all the engineering help they can get.

The final consequence can best be understood in terms of the Darwinian dictum that when a system is isolated and allowed to evolve of its own accord, it will inevitably mutate and deviate from the norm. Thus when school mathematics education is isolated from mathematicians, so is school mathematics itself, and, sure enough, the latter evolves into something that in large part no longer bears any resemblance to mathematics. Correct definitions are not given, or if given, they are not put to use (Milgram-Wu [18], Wu [23], [27] and [29]). The organic coherence of mathematics is

[8]This only tells half the story about mathematicians. See the comments near the end of this article.

no longer to be found (Wu [23]), or when "mathematical connections" are intentionally emphasized, such "connections" tend to be the trivial and obvious kind. Logical deduction becomes an afterthought; proofs, once relegated to the secondary school geometry course, were increasingly diluted until by now almost no proofs at all are found there, or anywhere else in the schools (Wu [26]). And so on. This development naturally brings down the quality of many aspects of mathematics education.

The absence of dialog between the two communities has led to many engineering errors in mathematics education, one of them being the unwelcome presence of mathematically incorrect test items in state and other standardized tests (Milgram [16] and [17]). The same kind of defective items also mar many teachers' credentialing tests (Askey [1] and [2]). A more subtle effect of the absence of mathematical input on assessment is the way test scores are routinely misinterpreted. The low test scores have been used to highlight students' dismal mathematical performance, but little or no thought is given to the possibility that they highlight not necessarily students' achievement (or lack thereof) but the pervasive damage done by defective curricular materials, or even the chronic lack of effective teaching. Such a possibility may not be obvious to anyone outside of mathematics, but to a mathematician, it does not take any research to confirm the fact that when students are taught incorrect mathematics, they learn incorrect mathematics. Garbage in, garbage out. If the incorrect mathematics subsequently shows up in students' test scores, how can we separate the errors due to the incorrect information students were given, from the errors due to students' own misconceptions? A more detailed examination of this idea in the narrow area of school algebra is given in Wu [31]. The need for mathematicians' participation in all phases of assessment is all too apparent.

The lack of collaboration between mathematicians and mathematics educators affects professional development as well. The issue of teacher quality is now openly acknowledged and serious discussions of the problem are beginning to be accepted in mathematics education (cf. Ma [15], and Conference Board of the Mathematical Sciences [6][9]). As a result of the inadequate mathematics instruction teachers receive in K-12, their knowledge of mathematics is, by and large, the product of the mathematics courses they take in college.[10] In very crude terms, the number of such required mathematics courses is too low, and in addition, these courses are taught either by mathematicians who are not in close consultation with teachers, and are unaware as to what is needed in the school classroom, or by mathematics educators who are not professional mathematicians. The former kind of course tends to be irrelevant to the classroom, and the latter kind tends to be mathematically shallow or incorrect. It is only natural that teachers coming out of such an environment turn out to be mathematically ill-prepared.

[9]Whatever reservations one may have concerning the details of its content, it is the fact that such a volume could be published under the auspices of a major scientific organization that is important.

[10]It may be useful to also take note of what may be called "the second order effect" of university instruction: teachers' knowledge of mathematics is also conditioned by their own K-12 experiences, but these teachers' teachers were themselves products of the mathematics courses they took in the university.

Similar woes persist in in-service professional development, thereby ensuring that teachers have little access to the mathematical knowledge they need for their profession. For example, the last decade has witnessed the appearance of case books consisting of actual records of lessons given by teachers.[11] The idea is to invite teachers to analyze these lessons, thereby sharpening their pedagogical sensibilities. In too many instances, however, blatant mathematical flaws in the cited cases are overlooked in the editors' commentaries. This raises the specter of bringing up a generation of teachers who are proficient in teaching school students incorrect mathematics. In this instance, it would appear that the need to respect mathematical integrity in mathematical engineering has been all but forgotten.

The most divisive outcome of the noncommunication between the two communities in the U.S. is undoubtedly the conflict engendered by the new (reform) curricula written in the past fifteen years. I take up this discussion last, because it brings us face to face with some subtle issues about mathematicians' participation in K-12 mathematics education. The prelude to the writing of these curricula is the unchecked degeneration in the mathematical integrity of the existing textbooks from major publishers over the period 1970–1990, a fact already alluded to above. This degeneration triggered the reform spearheaded by NCTM (National Council of Teachers of Mathematics [19]). Rightly or wrongly, the new curricula were written under the banner of the NCTM reform, and the manner in which some of the reform texts were imposed on public schools led eventually to the well-known Math Wars (Jackson [10]). The root of the discontent over these texts is the abundance of outright mathematical errors[12], as well as what research mathematicians perceived to be evidence of a lack of understanding of the mathematics. An example of the latter was the promotion of children's invented algorithms at the expense of the standard computation algorithms in the elementary mathematics curriculum. Although the promotion was partly an overreaction to the way the standard algorithms were often inflicted on school children with nary a word of explanation, it also reflected a lack of awareness of the central importance of the mathematical lessons conveyed by the reasoned teaching of these algorithms.

The "subtle issues" mentioned above stem from the fact that the writing of some of the new reform curricula actually had the participation of a few mathematicians. The first thing to note is that the latter are the rare exceptions to the general noncommunication between the mathematics and education communities. The noncommunication is real. At the same time, these exceptions seem to point to an apparent contradiction: How would I reconcile my critical stance toward these reform curricula with the principal recommendation of this article, namely, that mathematicians be equal partners with educators in the mathematics education enterprise? The answer is that there is no contradiction at all. The participation by mathematicians is, in general terms, a prerequisite to any hope of success in K-12 mathematics education, but in no way

[11] Let it be noted explcitly that I am discussing the case books in K-12 mathematics education only.

[12] These errors tend to be different from the earlier ones to be sure, but errors they are.

does it guarantee success. It is helpful in this context to recall similar discussions that routinely took place some eight years ago when some mathematicians first went public with the idea that mathematics teachers must have a solid content knowledge. The usual rejoinder at the time was that "knowing mathematics is not enough (to be a good teacher)". This is a common confusion that mistakes a necessary condition for a sufficient condition.[13] There is no quick fix for something as complex as mathematics education. Getting mathematicians to fully participate is only the beginning; the choice of the mathematicians and the hard work to follow will have a lot to say about the subsequent success or failure.

It is appropriate at this point to recall what was said at the beginning of the article about the power structure of mathematics education: thus far, educators get to make the decisions. Granting this fact, I should amplify a bit on the difficulties of choosing the right mathematicians for education work. Mathematicians have a range of background and experiences and, consequently, often have a range of opinions on matters of education as well. It is important that the range of these opinions be considered in all aspects of education. Many of the less happy incidents of the recent past in K-12 mathematics education were the result of choosing mathematicians of a particular persuasion. In addition, educators must make their own judgement on which among the mathematicians interested in K-12 are knowledgeable about K-12. Among the latter, some possess good judgment and leadership qualities while others don't. Educators must choose at each step. If there are algorithms for making the right choices, I don't happen to know them.

Every mathematician potentially has something to offer in K-12 mathematics education: even an occasional glance at textbooks to check for mathematical correctness can be very valuable. However, if mathematicians want to participate in serious educational work in K-12, what must they bring to the table? I believe the most important thing is the awareness that K-12 mathematics education is not a subset of mathematics, and that there is quite a bit to learn about the process of customization that distinguishes K-12 mathematics education from mathematics. In particular, much (if not most) of the mathematics they teach in the university cannot be brought straight to the school classroom (Wu [22]; Kilpatrick et al. [12], Chapter 10 and especially pp. 375–6), but that it must first go through the engineering process to make it suitable for use in schools. If I may use the example of fractions once again, mathematicians interested in making a contribution to K-12 may find it instructive to get to know the reason that something like "equivalence classes of ordered pairs of integers" is totally opaque to students around the age of twelve. They would also want to know the reason that students of that age nonetheless need a definition of a fraction which is as close to parts-of-a-whole as possible. They should also get to know the appropriate kind of mathematical reasoning for students in this age group, because they will ultimately be called upon to safeguard such reasoning in the curriculum and

[13] And need I point out, there are some who intentionally use this confusion to reject that mathematical content knowledge is important for teachers, or that getting mathematicians to participate in mathematics education is critical for its success.

assessment for these students.

Mathematicians may regard school mathematics as technically primitive (in the sense of skills), but they must take note of its conceptual sophistication (Jensen [11]; Wu [24], [25] and [30]; cf. also Aharoni [1]). Above all, they must know that school mathematics is anything but pedagogically trivial: There is absolutely nothing trivial about putting any material, no matter how simple, into a correct mathematical framework so that it may be profitably consumed by school students. Mathematicians who want to contribute to K-12 mathematics education have to be constantly on the alert to ensure that the minimum requirements of their profession – the orderly and logical progression of ideas, the internal cohesion of the subject, and the clarity and precision in the presentation of concepts, – are still met in mathematics education writings. This is no easy task. If mathematicians want to enter K-12 mathematics education as equal partners with educators, then it is incumbent upon them to uphold their end of the bargain by acquiring this kind of knowledge about mathematical engineering.

The concept of mathematics-education-as-mathematical-engineering does not suggest the creation of any new tools for the solution of the ongoing educational problems. What it does is to provide a usable intellectual framework for mathematics education as a discipline, one that clarifies the relationship between the mathematics and the education components, as well as the role of mathematicians in mathematics education. For example, it would likely lead to a better understanding of why the New Math became the disaster that it did. Most importantly, this concept lays bare the urgent need of the mathematical presence in every aspect of K-12 mathematics education, thereby providing a strong argument against the self-destructive policy of keeping mathematicians as outsiders in mathematics education. The chasm between mathematicians and educators must be bridged if our children are to be better served. I am cautiously optimistic[14] that there are enough people who want to rebuild this bridge (cf. Ball et al. [4]), all the more so because the indications are that the NCTM leadership is also moving in the same direction. I look forward to a future where mathematics education is the joint effort of mathematicians and educators.

**Acknowledgement.** I am first of all indebted to my colleague Norman E. Phillips for providing a critical piece of information about chemistry that got this article off the ground. The suggestion by Tony Gardiner to re-organize an earlier draft, and the penetrating comments on that draft by Helen Siedel, have left an indelible imprint on this article. Tom Parker, Ralph Raimi, and Patsy Wang-Iverson gave me very detailed corrections. David Klein also made corrections and alerted me to one of the references. In addition, the following members of the e-list mathed offered suggestions for improvement: R. A. Askey, R. Bisk, E. Dubinsky, U. Dudley, T. Foregger, T. Fortmann, K. Hoechsmann, R. Howe, W. McCallum, J. Roitman, M. Saul, D. Singer, A. Toom. Cathy Seeley and Skip Fennell also made similar suggestions.

It gives me pleasure to thank them all.

[14] In January of 2006.

## References


[1] Aharoni, R., What I Learned in Elementary School. American Educator, Fall, 2005; http://www.aft.org/pubs-reports/american_educator/issues/fall2005/aharoni.htm.

[2] Askey, R. A., Learning from assessment. In volume of the 2004 presentations in the Mathematical Sciences Research Institute, to appear.

[3] Askey, R. A., Mathematical content in the context of this panel. In *Proceedings of the Tenth International Congress on Mathematical Education* (ed. by Mogens Niss et al.), 2006.

[4] Ball, D. L., Ferrini-Mundy, J., Kilpatrick, J., Milgram, J. R., Schmid, W., Schaar, R., Reaching for common ground in K-12 mathematics education. *Notices Amer. Math. Soc.* **52** (2005), 1055–1058.

[5] Bass, H., Mathematics, mathematicians, and mathematics education. *Bull. Amer. Math. Soc.* **42** (2005), 417–430.

[6] Conference Board of the Mathematical Sciences. The Mathematical Education of Teachers, CBMS Issues in Mathematics Education 11, Amer. Math. Soc., Providence, RI, 2001.

[7] Education Week. A Purge at the French High Committee for Education (HCE). *Education Week*, November 27, 2005. http://www.educationnews.org/A-Purge-at-the-French-High-Committee-for-Education-HCE.htm.

[8] Ferrini-Mundy, J. Findell, B., The mathematics education of prospective teachers of secondary school mathematics: old assumptions, new challenges. In CUPM Discussion Papers about Mathematics and the Mathematical Sciences in 2010: What Should Students Know? Washington DC: Mathematical Association of America, Washington DC, 2001.

[9] Hill, H., Rowan, B., Ball, D. L., Effects of teachers' mathematical knowledge for teaching on student achievement, 2004. http://www-personal.umich.edu/dball/ BallSelectPapers TechnicalR.html

[10] Jackson, A., The Math Wars: California battles it out over mathematics education reform. *Notices Amer. Math. Soc.* **44** (1997), Part I, 695–702; Part II, 817–823.

[11] Jensen, G., Arithmetic for Teachers. Amer. Math. Soc., Providence, RI, 2003.

[12] Kilpatrick, J. Swafford, J. Findell, B., eds., Adding It Up. National Academy Press, Washington DC, 2001.

[13] Klein, D. et al., The state of State MATH Standards. Thomas B. Fordham Foundation, Washington DC, 2005. http://www.edexcellence.net/foundation/publication/publication.cfm?id=338.

[14] Lagemann, E. C., *An Elusive Science: The Troubling History of Education Research*. The University of Chicago Press, Chicago, London 2000.

[15] Ma, L., *Knowing and Teaching Elementary Mathematics*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[16] Milgram, R. J., Problem solving and problem solving Models for K-12: Preliminary Considerations. 2002; http://math.stanford.edu/ftp/milgram/discussion-of-well-posed-problems.pdf.

[17] Milgram, R. J., Pattern recognition problems in K-12. 2003; http://math.stanford.edu/ftp/milgram/pattern-problems.pdf.

[18] Milgram, R. J. and Wu, H., Intervention program. 2005; http://math.berkeley.edu/~wu/.

[19] National Council of Teachers of Mathematics. Curriculum and Evaluation Standards for School Mathematics. National Council of Teachers of Mathematics, Reston, VA, 1989.

[20] Shulman, Lee, Those who understand: Knowledge growth in teaching. *Educational Researcher* **15** (1986), 4–14.

[21] Washington Post. An Open Letter to United States Secretary of Education, Richard Riley. November 18, 1999; http://mathematicallycorrect.com/nation.htm.

[22] Wu, H., On the education of mathematics teachers (formerly entitled: On the training of mathematics teachers). 1997; http://math.berkeley.edu/~wu/.

[23] Wu, H., What is so difficult about the preparation of mathematics teachers? 2001; http://math.berkeley.edu/~wu/.

[24] Wu, H., Chapter 1: Whole Numbers (Draft). 2001; http://math.berkeley.edu/~wu/.

[25] Wu, H., Chapter 2: Fractions (Draft). 2001; http://math.berkeley.edu/~wu/.

[26] Wu, H., Geometry: Our Cultural Heritage – A book review. *Notices Amer. Math. Soc.* **51** (2004), 529–537.

[27] Wu, H., Key mathematical ideas in grades 5-8. 2005; http://math.berkeley.edu/~wu/.

[28] Wu, H., Must content dictate pedagogy in mathematics education? 2005; http://math.berkeley.edu/~wu/.

[29] Wu, H., Professional development: The hard work of learning Mathematics. 2005; http://math.berkeley.edu/~wu/.

[30] Wu, H., Introduction to School Algebra (Draft). 2005; http://math.berkeley.edu/~wu/.

[31] Wu, H., Assessment in school algebra. 2006, to appear.

## The role of mathematicians in K-12 education: a personal perspective

by *Shiu-Yuen Cheng*

This draft was written after I read Prof. H. Wu's draft on "How mathematicians can contribute to K-12 mathematics education". I therefore have the advantage of adopting the same terms and scope of discussions in writing this draft. For example, I will be using Prof. Wu's definition of the word "mathematician" to mean "research mathematicians". Also, I am impressed and I agree with Prof. H. Wu's philosophical idea of regarding mathematics education as mathematical engineering. I will in the following outline the roles that mathematicians can play for the enhancement of K-12 education. Frequently, I will come back to Prof. Wu's idea of mathematical engineering so that we can do a good job.

The main factors for providing an effective mathematics education are curriculum design, teacher competence and assessment methods. Among these three factors, I think the most important one is teacher competence. I think this is the factor that mathematicians can contribute the most. The processes of designing the curriculum and assessment mechanism vary from place to place and are greatly influenced by the local bureaucratic and political system. In most places, mathematicians do not

get to play much of a role in the design of curriculum and the assessment mechanism. However, this does not mean that we should fold our hand and watch on the sideline. We should always engage in these two factors and make our contributions whenever possible. On the other hand, any curriculum design or assessment method in mathematics would need or welcome mathematicians' stamp of approval. Mathematicians will definitely be involved but we have to vigilantly and patiently engage in the process.

In Hong Kong, the relation between educators and mathematicians is much better than in the US. However, the educators do not call the shot. Instead, the government officials set the agenda and play the most influential role. On paper, it does not seem so because things are supposedly done through committees consisting of teachers, principals, educators and mathematicians. The government officials serve as secretariats in the committees. As the committee members are all busy people taking time off from their work to attend the committee meetings, the secretariat then get to draft all the papers and the agenda. They naturally become most influential. Moreover, as a consequence of the composition of the committees mathematicians are minority. To make things worse, they have few allies. The teachers, school principals in the committee usually assume the mathematicians in the committee have a secret agenda to tailor the curriculum for attracting students to be mathematics majors. Additionally, the educators talk the language of education officials. Their inputs are more helpful to the education officials in filling the reports with popular education jargons. It is then natural that mathematicians' views usually do not prevail. Instead, some compromise can usually be reached if mathematicians engage in the process.

I believe that teacher competence is most important factor as teachers are at the frontline implementing the curriculum and delivering the mathematics education. No matter how hard we work, usually the curriculum and assessment mechanism are far from perfect. A competent teacher can exercise discretions to compensate the inconsistencies and incompleteness of the curriculum and make them work. On the other hand, a teacher who has little confidence and competence in subject knowledge can easily turn a well-designed curriculum or assessment mechanism into disasters. In the area of teacher competence, mathematicians can contribute in two main areas: the university curriculum for mathematics teacher program, and courses and workshops for in-service mathematics teachers. Mathematicians can play major roles in these two areas and can get more colleagues to participate. However, we usually do not pay much attention or do not do the right thing. It was pointed out by Prof. H. Wu and many others that the university curriculums for mathematics majors do not serve the purpose of providing the necessary understanding to be a competent mathematics teacher. The main reason is that the curriculum for mathematics majors is designed with an aim to train research mathematicians. As for courses and workshops for in-service teachers, we need more colleagues to participate and contribute. The ball is in our court but so far we have not made the right play.

In Hong Kong, about fifteen to twenty percent of mathematics graduate become mathematics teachers. This is not a small percentage and is in fact higher than the

percentage of students going for postgraduate study in mathematics. However, the curriculum for mathematics major does not offer much help for those who will pursue the career of a mathematics teacher. Mainly, most mathematicians do not see the necessity of designing and offering some new courses to provide a profound understanding of school mathematics. It is assumed that our courses in abstract algebra, analysis and geometry will do the job and hence nothing needs to be done. As for courses and workshops for in-service teachers, the sad thing is that Hong Kong government does not provide much of this kind of opportunity. Mathematicians have to shoulder this task on a volunteer basis. There are mathematicians willing to contribute but in order to make it sustainable we need to convince the government and the mathematics community the importance of providing courses and workshops for deepening teachers' understanding of the subject knowledge. To do this effectively, we need to communicate well to the community about the concerns of mathematicians about mathematics education. In many places, people are alarmed because "Johnny can't add". In Hong Kong, Johnny can add! In fact, Johnny can do fractions and decimals quite well. In many international studies about the mathematics attainment of school students, Hong Kong routinely occupies one of the top positions. On paper, we should congratulate ourselves and should not even attempt to touch the system as things are not broken. However, anyone in the university or familiar with the situation of the Hong Kong school system knows that the Hong Kong mathematics education is far from achieving the goals. We are able to train our students to do arithmetic and some simple algebra at a tremendous cost. In the process, we suffocate students' creativity and motivation for learning. So far we have not been successful in documenting and communicating our concerns to the Hong Kong public and the government. Mathematics education is then getting little resource from the government as it is doing quite well comparing to our language education. I believe communicating effectively to the public about our views for enhancing mathematics education is crucial and should fall into one of the sub-areas of Prof. Wu's framework of Mathematics Engineering. The banner of Mathematics Engineering is useful for setting a clear goal and rallying support of our fellow mathematicians to contribute to mathematics education.

## The role of mathematicians in K-12 mathematics education

by *U. Kirchgraber and K. Osterwalder*

We begin with a few remarks on the Swiss educational system and on teaching and learning of mathematics in Switzerland. Then we focus on teacher training in Mathematics at the Swiss Federal Institute of Technology (ETH). Finally we sketch an answer to the question posed to the panel.

In international comparative studies like TIMS and PISA Swiss students have demonstrated reasonable achievements in Mathematics. Without overestimating such

results[15] one may wonder whether some specific features of the Swiss educational system might be responsible for this relative success and which measures could serve a further improvement of the results. As we will see some of the possible explanations are related to the topic of the panel.

It is well known Switzerland is not rich in natural resources. This is usually claimed to be one of the major reasons why education is quite highly valued in this country, with a number of important implications: the vast majority of schools are public, teachers on all levels are well trained, the profession of teacher is quite respected, teachers (on all levels) are well paid, schools are well equipped, school buildings are kept in good shape.

The Swiss educational system leaves considerable freedom to teachers on all levels and in particular in upper secondary school, to which we will refer to as Gymnasium[16] level[17]. Gymnasium teachers in general and Gymnasium Mathematics teachers in particular have to follow a certain core curriculum. Yet beyond this guide line they are fairly free to include additional topics, there are hardly any restrictions concerning the type of pedagogy adopted, and teachers are even quite free as to the number and type of tests and examinations they will administer. As to the final examination, in many schools every single mathematics teacher is free to assign a selection of, say, 4-8 problems of his or her choice and depending on the topics he or she has covered in class to his or her students on which they will work during 4 hours. Eventually he or she will correct and grade these works.

ETH offers Gymnasium Teacher Training Programs in the following fields: Biology, Chemistry, Earth Sciences, Mathematics, Physics and Sports. In the following we discuss some features of the Gymnasium Mathematics Teacher Training Program (GMTTP).

A prerequisite for completing the GMTTP is a Master's Degree in Mathematics, though the students are permitted to start with the GMTTP in the third year of the Bachelor's program. Average duration time for completing the GMTTP is six months[18], if studied full time.

The fact that Swiss Gymnasium Mathematics teachers must hold a Master's Degree in Mathematics has – we suppose – far reaching professional and psychological consequences. Having completed a Master's program has at least two implications which, we think, are important for a future Mathematics teacher: a) During the first two years of studies Mathematics students encounter many topics they have seen before yet dealt with in way that is qualitatively very different from what they had

[15] Compared to the host of highly sophisticated tools to measure many quantities in the Sciences and in particular in Physics that have evolved since the time of Gallilei, measuring effectiveness of teaching and learning and similar variables is probably still in its infancy, yet is a fascinating and challenging enterprise.

[16] There are some 150 Gymnasia in Switzerland, every year some 15000 students graduate from the Gymnasia, between 1800 and 2300 enroll at ETH.

[17] This corresponds to grades 9–12.

[18] Starting in fall 2006 federal requirements request studies twice as long. ETH's GMTTP will be extended accordingly and will be renamed as "Master of Advanced Studies in Secondary and Higher Education in Mathematics".

experienced previously. b) In the third year of the Bachelor's and during the Master's program they are exposed to advanced fields, an indispensable experience for gaining a faithful picture of what Mathematics is about.

Based on four and a half years of studies these teacher students can at least potentially be expected to dispose of a degree of mathematical expertise and mathematical maturity which is covered by the terms "content knowledge" and/or "deep understanding" in the Mathematics Education research literature.

As to some practical implications: Being trained as full fledged mathematician a Gymnasium Mathematics teacher may leave school after a few years and start a career in Industry or elsewhere, or vice versa. Therefore requiring a Master's Degree as a prerequisite for teacher training has the benefit of not excluding Gymnasium Mathematics teachers in an early stage of their professional development from the full scale of professional opportunities offered to mathematicians nowadays.

The GMTTP includes courses in the Educational sciences, in Mathematics Education (Didactics of Mathematics), (a small amount of) guided teaching practice, and a fourth component, called Specialized Mathematics courses with an Educational Focus. It is of utmost importance that these components are excellently tuned and multiply intertwined. Moreover they should be accompanied by plenty of student activities[19].

Since the late eighties the Educational course at ETH was designed and continuously updated. The basic concept was to make available both to the Science and Mathematics Educators as well as to their teacher students research grounded results from areas such as psychology, the cognitive sciences, etc. Over the years quite a number of teaching techniques and teaching methods were implemented and probed since they are known – on the basis of meta analyses[20] – to enhance learning[21]. Guided Learning programs[22] for instance are self-contained study materials for pupils covering a learning unit of some 3–30 lessons with the following features: Precisely defined prerequisites, well structured and comprehensibly written explanations, explicitly stated learning goals, adjunct questions and their answers, learning aids, chapter tests to fulfill the so-called Mastery Learning Principle. According to Kulik, Kulik and Bangert-Drowns[23] the effect size of this type of teach ware is of the order of 0.5 in Mathematics and of the order of 0.6 in the Sciences. A few examples of Guided Learning Programs in Mathematics and the Sciences (in German) can be found on www.educeth.ethz.ch[24].

[19] As research has shown, just attending lectures has little impact on future teaching.

[20] See for instance Fraser B. J., Walberg H. J., Welch W. W., Hattie J. A., Syntheses of Educational Productivity Research. *International J. of Educational Research* **11** (1987), 145–252; Walberg, H. J., Productive Teaching and Instruction: Assessing the Knowledge Base. University of Illinois at Chicago, School of Education, 1988, 18 p, mimeographed.

[21] I.e. they have noteworthy effect sizes.

[22] In German: Leitprogramme.

[23] Kulik F. S., Kulik J. A., Bangert-Drowns R. L., Effectiveness of Mastery Learning Programs: A Meta-Analysis. *Review of Educational Research* **60** (1990), 265–299.

[24] EducETH, a service of ETH to the Public, is ETH's educational server providing teaching materials primarily for upper secondary schools.

In the Mathematics Education courses the thrust is on domain specific aspects of the teaching and learning enterprise. Of course, goals, standards, competencies to be achieved are discussed, subject matter analysis with diverse approaches to selected topics is of central concern and textbooks[25] are analyzed. The teacher students are exposed to Mathematics Education research concepts that prove useful to explain certain phenomena, for instance Talls's and Vinner's distinction between concept definition and concept image which helps to understand students' misconceptions of the notion of function. Topics like the "Expert Blind Spot"-Hypothesis[26], or the influence of teacher's pedagogical content beliefs on learning outcomes[27], and many others are treated. Videos and their transcriptions are analyzed to provide insight into the unpredictability and fragility of learning processes, among other things.

We now turn to the Specialized Mathematics courses with an Educational Focus already mentioned before. It is by now generally accepted that transfer achievements quite often do not emerge automatically. F. Weinert, summarizing years of research at the Max Planck Institut für psychologische Forschung in Munich, explains it roughly speaking as follows. Knowledge a learner acquires systematically – for instance in Mathematics courses as they are usually organized – is likely to be structured and organized in the learners brain in a way not easily retrievable, amenable if the learner is put in a problem situation in which he/she should apply this body of knowledge. Thus, knowledge, which is available in principle, remains inert and unused, though it would be useful and even necessary to handle a certain situation. Weinert's conclusion: To build up an intelligent, flexibly applicable knowledge base the learner needs both, systematic as well as situated learning.

The Specialized Mathematics courses with an Educational Focus take place in the third year (in Switzerland: the last year) of the Bachelor's and during the Master's program. They are open to all students in the Bsc/Msc Mathematics program, but are compulsory for candidates in the GMTTP. These courses with an Educational focus were installed many years ago at ETH (long before there were courses on Mathematics Education!) and have their origin in lectures given by Felix Klein early in the 20th century in Göttingen under the title "Elementarmathematik vom höheren Standpunkt" (elementary mathematics from an advanced point of view) and directed to future Gymnasium Mathematics teacher.

The Specialized Mathematics courses with an Educational Focus serve several goals. Very much in the spirit of Klein's concept they attempt to narrow the gap between Gymnasium Mathematics and University Mathematics. Take a core topic in Mathematics, present at all levels: equations. Linear equations are a topic in grade 9,

[25] After what has been said earlier in this paper it will not come as a big surprise for the reader that Swiss Gymnasium Mathematics teacher are not obliged to use any particular text books. Many in fact do not use a textbook at all but use a variety of sources to assemble handouts, etc., for their students.

[26] See: M. J. Nathan, A. Petrosino: Expert Blind Spot among Pre-service Teachers. *Amer. Educ. Res. J.* **40** (2003), 905–928.

[27] See: Staub, F. C. and Stern, E., The Nature of teachers' Pedagogical Content Beliefs Matters for Students' Achievement Gains. *J. Educational Psychology* **93** (2002), 344–355.

in a course on Linear Algebra in the first year of the Bachelor's program, as well as in specialized courses on Numerical Linear Algebra: Relate the various aspects and draw conclusions for the future teaching of linear equations in grade 9[28]!

Pupils usually encounter nonlinear equations first in connection with the quadratic equation. Most emphasis is usually put on reducing a general quadratic equation[29] to a "purely quadratic" equation[30]. It is of course a marvelous discovery that arbitrary quadratic equations can be reduced to purely quadratic ones. It is a challenging design task for teacher students to compose a series of assignments that guides pupils to discover this phenomenon by themselves.

Yet from a more general point of view the question of solving purely quadratic equations is even more intriguing. One encounters a pattern that is prevalent in (University) Mathematics: Equations are not always solvable. More often than not mathematicians have to invent a setting such that the equation becomes solvable: Loosely speaking – mathematicians make equations solvable! A second such instance comes up when complex numbers are invented to make all quadratic equations solvable with the totally unexpected benefit that in this setting all polynomial equations have solutions.

The Fundamental Theorem of Algebra brings up another very interesting phenomenon: we may be able to prove that an equation has a solution and even that this solution is unique without being able to compute the solution. If complex numbers are treated at the Gymnasium[31] level an intuitive proof of the Fundamental Theorem of Algebra can be offered to Gymnasium students. In a Specialized Mathematics course with an Educational Focus dedicated to equations, teacher students would not only design a learning unit for pupils around such a heuristic proof, but learn in addition how such a proof is made rigorous (not an easy task!), topics like Rouché's Theorem would have to be discussed and an introduction to the Brouwer and Leray–Schauder Degree theory with applications to periodic solutions of differential equations would allow for a glance of the breadth of the field.

Another aspect to which Specialized Mathematics courses with an Educational Focus can contribute concerns curricular development. School Mathematics curricula are often blamed for covering material only that was invented centuries ago. Of course, most subjects that are hot research topics in Mathematics are far too remote and far too specialized for being accessible at the Gymnasium level. Yet there are marvelous exceptions: The Diffie-Hellman Key Exchange and RSA Cryptography, invented in

[28] Maybe you conclude that the 9th grade program on linear equation should contain a modest introduction to Computerized Tomography, as we did, see the Leitprogramm entitled "Gleichungen" at www.educeth.ethz.ch. Maybe you conclude that the program, in addition to Gaussian elimination, should include a homeopathic introduction to solving linear equations by iteration (a topic you might touch on again, when you treat Banach fixed point iteration in one dimension in connection with Kepler's equation). Maybe you conclude that the question of what it means that two linear systems of equations are equivalent, and how one obtains equivalent systems from a given one, deserves to become (a small?) Mathematics Education research project.

[29] I.e. one including a linear term with respect to the unknown.

[30] I.e. one in which the linear term is absent.

[31] Gymnasia have various different profiles in Switzerland. Some concentrate on Mathematics and Physics. There complex numbers are treated.

the late seventies, are well suited to give 10th graders an idea of the mathematical enterprise[32].

Euler buckling is probably the earliest example of a bifurcation problem. It was only during the last two or three decades, however, that bifurcation theory became a systematically developed branch of analysis. In a Specialized Mathematics course with an Educational Focus dedicated to an introduction to bifurcation theory as background it is well possible that teacher students adapt some of the material for Gymnasium students, hereby heavily drawing on a classic school subject: the study of the geometrical properties and graphing of functions defined by simple expressions.

Finally we mention ill-posed inverse problems. This again is a relatively new field of Applied Mathematics. It is of great theoretical and practical interest, the way ill-posed inverse problems are treated mathematically is surprising and they lend themselves outstandingly for treatments on various different levels[33].

We expect that Specialized Mathematics courses with an Educational Focus deepen the teacher students' mathematical expertise, that they strengthen the link between University and Gymnasium level Mathematics, that they contribute to develop the secondary school Mathematics curricula, that they support the prospective teachers to teach Mathematics at the same time more mathematically and in such a way that their students can learn to value Mathematics as a human activity and for its significance in our world.

What is eventually the role of mathematicians in K-12 in our system?

Research mathematician can and do contribute in a number of ways. Via the Bachelor's and Master's program they shape lastingly the knowledge base, the picture and the skills our teacher students develop. They can substantially contribute to the Specialized Mathematics courses with an Educational Focus. They can contribute to the design of substantial teaching units.

In Mathematics Education research highly interesting developments have just begun. We mentioned the paper by Staub and Stern entitled: "The Nature of teachers' Pedagogical Content Beliefs Matters for Students' Achievement Gains." Another paper in the same realm is by Hill, Rowan and Ball[34]. It is entitled: "Effects of teachers' mathematical knowledge for teaching on student achievement." These papers provide results on primary school teachers and primary school pupils. We certainly need many more results on "what has which impact on student learning" and very much so in higher grades. In fact, very little seems to be known as to the upper Gymnasium level.

[32]Here are some aspects: Fermat's (little) Theorem on which RSA cryptography is based, though elementary, is far from being obvious. The way it is proven illustrates the power of mathematical ideas. 350 years after its discovery it became the key ingredient to affirmatively answer a questions one hardly dares to ask: Is it possible that two persons, who cannot communicate but publicly, can agree on keys which permit them to exchange messages which cannot be decoded except by the person who is entitled to read the message?

[33]See Kirchgraber, U., Kirsch, A., Stoffer, D. : Schlecht gestellte Probleme – oder wenn das Ungenaue genauer ist. *Math. Semesterber.* **51** (2004), 175–2005.

[34]Hill, H. C., Rowan, B., Ball, D. L.: Effects of teachers' mathematical knowledge for teaching on student achievement. *Amer. Educ. Res. J.* **42** (2005), 371–406.

Earlier we noted that educational measurement techniques are probably still in their infancy. How can we suitably measure mathematical achievements, teachers' pedagogical content knowledge, the nature of teachers' pedagogical content beliefs, and many more variables of this type? This is certainly a field, where research mathematician can and should contribute.

Research mathematicians in Switzerland are welcome as members in school boards, and/or as experts in the final examinations at Gymnasium Schools. Research mathematicians are welcome to offer lectures to in-service teachers or to participate in study weeks for Gymnasium students, or to visit schools and give talks.

To summarize: The main contribution of research mathematicians to the second half of K-12 is to train Mathematics teachers as knowledgeable mathematicians and to develop with them methods to narrow the gap between "Gymnasium Mathematics" and University Mathematics. Other possible contributions are manifold, crucial and indispensable.

Ateneo de Manila University, Quezon City 1108, Philippines
E-mail: bnebres@ateneo.edu

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China
E-mail: macheng@ust.hk

ETH-Zentrum, Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland
E-mail: osterwalder@sl.ethz.ch, kirchgra@math.ethz.ch

Department of Mathematics, University of California, Berkeley, CA 94720-3840, U.S.A.
E-mail: wu@math.berkeley.edu

# On the origins of Hilbert's sixth problem: physics and the empiricist approach to axiomatization

Leo Corry

**Abstract.** The sixth of Hilbert's famous 1900 list of twenty-three problems is a programmatic call for the axiomatization of physical sciences. Contrary to a prevalent view this problem was naturally rooted at the core of Hilbert's conception of what axiomatization is all about. The axiomatic method embodied in his work on geometry at the turn of the twentieth-century originated in a preoccupation with foundational questions related with empirical science, including geometry and other physical disciplines at a similar level. From all the problems in the list, the sixth is the only one that continually engaged his efforts over a very long period, at least between 1894 and 1932.

**Mathematics Subject Classification (2000).** Primary 01A60; Secondary 03-03, 70-03, 83-03.

**Keywords.** David Hilbert, axiomatization, physics.

## 1. Introduction

Of the many important and brilliant plenary talks delivered in ICMs ever since the inception of this institution in 1897 in Zurich, none has so frequently been quoted and, possibly, none has had the kind of pervasive influence, as the one delivered by David Hilbert in 1900 at the second ICM in Paris, under the title of "Mathematical Problems". Rather than summarizing the state of the art in a central branch of mathematics, Hilbert attempted to "lift the veil" and peer into the development of mathematics in the century that was about to begin. He chose to present a list of twenty-three problems that in his opinion would and should occupy the efforts of mathematicians in the years to come. This famous list has been an object of mathematical and historical interest ever since.

The sixth problem of the list deals with the axiomatization of physics. It was suggested to Hilbert by his own recent research on the foundations of geometry. He proposed "to treat in the same manner, by means of axioms, those physical sciences in which mathematics plays an important part." This problem differs from most others on Hilbert's list in essential ways, and its inclusion has been the object of noticeable reaction from mathematicians and historians who have discussed it throughout the years. Thus, in reports occasionally written about the current state of research on the twenty-three problems, the special status of the sixth problem is readily visible: not only has it been difficult to decide to what extent the problem was actually solved (or not), but one gets the impression that, of all the problems on the list, this one received

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

the least attention from mathematicians throughout the century and that relatively little effort was directed at solving it ([11], [25]).

Many a historical account simply dismissed the sixth problem as a slip on Hilbert's side, as a curiosity, and as an artificial addition to what would otherwise appear as an organically conceived list, naturally connected to his broad range of mathematical interests (e.g., [26], p. 159). In fact, this is how Hilbert's interest in physical topics in general as well as his few, well-known incursions into physical problems have been traditionally seen. According to this view, these are seen as sporadic incursions into foreign territory, mainly for the purposes of finding some new applications to what would otherwise be purely mathematically motivated ideas. This is the case, for instance, with Hilbert's solution of the Boltzmann equation in kinetic theory of gases in 1912. Starting in 1902, most of Hilbert's mathematical energies had been focused on research related with the theory of linear integral equations, and his solution of the Boltzmann equation could thus be seen as no more than an application of the techniques developed as part of that theory to a particular situation, the physical background of which would be of no direct interest to Hilbert. An account in this spirit appears in Stephen G. Brush's authoritative book on the development of kinetic theory, according to which:

> When Hilbert decided to include a chapter on kinetic theory in his treatise on integral equations, it does not appear that he had any particular interest in the physical problems associated with gases. He did not try to make any detailed calculations of gas properties, and did not discuss the basic issues such as the nature of irreversibility and the validity of mechanical interpretations which had exercised the mathematician Ernst Zermelo in his debate with Boltzmann in 1896–97. A few years later, when Hilbert presented his views on the contemporary problems of physics, he did not even mention kinetic theory. We must therefore conclude that he was simply looking for another possible application of his mathematical theories, and when he had succeeded in finding and characterizing a special class of solutions (later called "normal") …his interest in the Boltzmann equation and in kinetic theory was exhausted. ([4], p. 448)

A further important physical context where Hilbert's appeared prominently concerns the formulation of the gravitational field-equations of the general theory of relativity (GTR). On November 20, 1915, Hilbert presented to the Royal Scientific Society in Göttingen his version of the equations, in the framework of what he saw as an axiomatically formulated foundation for the whole of physics. During that same month of November, Einstein had been struggling with the final stages of his own effort to formulate the generally covariant equations that lie at the heart of GTR. He presented three different versions at the weekly meetings of the Prussian Academy of Sciences in Berlin, before attaining his final version, on November 25, that is, five days *after* Hilbert had presented his own version.

Einstein had visited Göttingen in the summer of 1915 to lecture on his theory and on the difficulties currently encountered in his work. Hilbert was then in the audience and Einstein was greatly impressed by him. Earlier accounts of Hilbert's involvement with problems associated with GTR had in general traced it back to this visit of Einstein or, at the earliest, to the years immediately preceding it. As in the case of kinetic theory, this contribution of Hilbert was often seen as a more or less furtive incursion into physics, aimed at illustrating the power and the scope of validity of the "axiomatic method" and as a test of Hilbert's mathematical abilities while trying to "jump onto the bandwagon of success" of Einstein's theory.

In biographical accounts of Hilbert, his lively interest in physics has never been overlooked, to be sure, but it mostly has been presented as strictly circumscribed in time and scope. Thus for instance, in his obituary of Hilbert, Hermann Weyl ([24], p. 619) asserted that Hilbert's work comprised five separate, and clearly discernible main periods: (1) Theory of invariants (1885–1893); (2) Theory of algebraic number fields (1893–1898); (3) Foundations, (a) of geometry (1898–1902), (b) of mathematics in general (1922–1930); (4) Integral equations (1902–1912); (5) Physics (1910–1922). Weyl's account implies that the passage from any of these fields to the next was always clear-cut and irreversible, and a cursory examination of Hilbert's published works may confirm this impression. But as Weyl himself probably knew better than many, the list of Hilbert's publications provides only a partial, rather one-sided perspective of his intellectual horizons, and this is particularly the case when it comes to his activities related to physics.

Recent historical research has brought to light a very different picture of Hilbert's involvement with physics, and in particular of the real, truly central place of the ideas embodied in the sixth problem within the general edifice of Hilbert's scientific outlook. Hilbert's involvement with physical issues spanned most of his active scientific life, and the essence of his mathematical conceptions cannot be understood without reference to that involvement. More importantly, the famous "axiomatic approach" that came to be identified with Hilbert's mathematical achievements and with his pervasive influence on twentieth-century mathematics is totally misunderstood if it is not seen, in the first place, as connected with his physical interests. Under this perspective, the involvement with kinetic theory and GTR are seen as a natural outgrowth of the development of Hilbert's world of ideas, and by no means as sporadic, isolated incursions into unknown territories. Moreover, contrary to a commonly held view, the sixth problem is the only one in the entire list of 1900 that refers to an idea that continually engaged the active attention of Hilbert for a very long period of time, at least between 1894 and 1932 ([5]).

The key to a balanced understanding of the role of physics within Hilbert's intellectual horizon is found not so much in his publications, as it is in the complex academic network of personal interactions and diverse activities that he was continually part of. Especially worthy of attention is his teaching, first at Königsberg and – more importantly – after 1895 at Göttingen. At the mathematical institute established by Felix Klein, Hilbert became the leader of a unique scientific center that brought

together a gallery of world-class researchers in mathematics and physics. One cannot exaggerate the significance of the influence exerted by Hilbert's thought and personality on all who came out of this institution. More often than not, these lectures were far from systematic and organized presentations of well-known results and established theories. Rather, Hilbert often used his lectures as a public stage where he could explore new ideas and think aloud about the issues that occupied his mind at any point in time. In a lecture held in commemorating his seventieth birthday, Hilbert vividly recalled how these lectures provided important occasions for the free exploration of yet untried ideas. He thus said:

> The closest conceivable connection between research and teaching became a decisive feature of my mathematical activity. The interchange of scientific ideas, the communication of what one found by himself and the elaboration of what one had heard, was from my early years at Königsberg a pivotal aspect of my scientific work. …In my lectures, and above all in the seminars, my guiding principle was not to present material in a standard and as smooth as possible way, just to help the student keep clean and ordered notebooks. Above all, I always tried to illuminate the problems and difficulties and to offer a bridge leading to currently open questions. It often happened that in the course of a semester the program of an advanced lecture was completely changed, because I wanted to discuss issues in which I was currently involved as a researcher and which had not yet by any means attained their definite formulation. ([16], p. 79)

The collection of Hilbert's lecture notes offers an invaluable source of information for anyone interested in understanding his scientific horizon and contributions.

## 2. Axiomatics and formalism

A main obstacle in historically understanding the significance of the sixth problem has been the widespread image of Hilbert as the champion of formalism in modern mathematics. The traditional association of Hilbert's name with the term "formalism" has often proved to be misleading, since the term can be understood in two completely different senses that are sometimes conflated. One sense refers to the so-called "Hilbert program" that occupied much of Hilbert's efforts from abut 1920. Although involving significant philosophical motivations, at the focus of this program stood a very specific, technical mathematical problem, namely, the attempt to prove the consistency of arithmetic with strictly finitist arguments. The point of view embodied in the program was eventually called the "formalist" approach to the foundations of mathematics, and it gained much resonance when it became a main contender in the so-called "foundational crisis" in mathematics early in the twentieth century.

Even though Hilbert himself did not use the term "formalism" in this context,

associating his name with term conceived in this narrow sense seems to be essentially justified. It is misleading, however, to extend the term "Hilbert program" – and the concomitant idea of formalism – to refer to Hilbert's overall conception of the essence of mathematics. Indeed, a second meaning of the term formalism refers to a general attitude towards the practice of mathematics and the understanding of the essence of mathematical knowledge that gained widespread acceptance in the twentieth century, especially under the aegis of the Bourbaki group. Jean Dieudonné, for instance, explained what he saw as the essence of Hilbert's *mathematical* conceptions in a well-known text where he referred to the analogy with a game of chess. In the latter, he said, one does not speak about truths but rather about following correctly a set of stipulated rules. If we translate this into mathematics we obtain the putative, "formalist" conception often attributed to Hilbert ([6], p. 551): "mathematics becomes a *game*, whose pieces are graphical *signs* that are distinguished from one another by their form."

Understanding the historical roots and development of the sixth problem goes hand in hand with an understanding of Hilbert's overall conception of mathematics as being far removed from Dieudonné's chess-game metaphor. It also comprises a clear separation between the "Hilbert program" for the foundations of arithmetic, on the one hand, and Hilbert's lifetime research program for mathematics and physics and its variations throughout the years, on the other hand. In this regard, and even before one starts to look carefully at Hilbert's mathematical ideas and practice throughout his career, it is illustrative to look at a quotation from around 1919 – the time when Hilbert began to work out the finitist program for the foundations of arithmetic in collaboration with Paul Bernays – that expounds a view diametrically opposed to that attributed to him many years later by Dieudonné, and that is rather widespread even today. Thus Hilbert said:

> We are not speaking here of arbitrariness in any sense. Mathematics is not like a game whose tasks are determined by arbitrarily stipulated rules. Rather, it is a conceptual system possessing internal necessity that can only be so and by no means otherwise. ([16], p. 14)

The misleading conflation of the formalist aspect of the "Hilbert program" with Hilbert's overall views about mathematics and its relationship with physics is also closely related with a widespread, retrospective misreading of his early work on the foundations of geometry in purely formalist terms. However, the centrality attributed by Hilbert to the axiomatic method in mathematics and in science is strongly connected with thoroughgoing empiricist conceptions, that continually increased in strength as he went on to delve into ever new physical disciplines, and that reached a peek in 1915–17, the time of his most intense participation in research associated with GTR.

The axiomatic approach was for Hilbert, above all, a tool for retrospectively investigating the logical structure of *well-established and elaborated* scientific theories, and the possible difficulties encountered in their study, and never the starting point for

the creation of new fields of enquiry. The role that Hilbert envisaged for the axiomatic analysis of theories is succinctly summarized in the following quotation taken from a course on the axiomatic method taught in 1905. Hilbert thus said:

> The edifice of science is not raised like a dwelling, in which the foundations are first firmly laid and only then one proceeds to construct and to enlarge the rooms. Science prefers to secure as soon as possible comfortable spaces to wander around and only subsequently, when signs appear here and there that the loose foundations are not able to sustain the expansion of the rooms, it sets about supporting and fortifying them. This is not a weakness, but rather the right and healthy path of development. ([5], p. 127)

## 3. Roots and early stages

Physics and mathematics were inextricably interconnected in Hilbert's scientific horizon ever since his early years as a young student in his native city of Königsberg, where he completed his doctorate in 1885 and continued to teach until 1895. Hilbert's dissertation and all of his early published work dealt with the theory of algebraic invariants. Subsequently he moved to the theory of algebraic number fields. But his student notebooks bear witness to a lively interest in, and a systematic study of, an astounding breadth of topics in both mathematics and physics. Particularly illuminating is a notebook that records his involvement as a student with the *Lehrbuch der Experimentalphysik* by Adolph Wüllner (1870). This was one of many textbooks at the time that systematically pursued the explicit reduction of all physical phenomena (particularly the theories of heat and light, magnetism and electricity) to mechanics, an approach that underlies all of Hilbert's early involvement with physics, and that he abandoned in favor of electrodynamical reductionism only after 1912.

In the intimate atmosphere of this small university, the student Hilbert participated in a weekly seminar organized under the initiative of Ferdinand Lindemann – who was also Hilbert's doctoral advisor – that was also attended by his good friends Adolf Hurwitz and Hermann Minkowski, by the two local physicist, Woldemar Voigt and Paul Volkmann, and by another fellow student Emil Wiechert, who would also become Hilbert's colleague in Göttingen and the world's leading geophysicist. The participants discussed recent research in all of branches of mathematics and physics, with special emphasis on hydrodynamics and electrodynamics, two topics of common interest for Hilbert and Minkowski throughout their careers. From very early on, fundamental methodological questions began to surface as part of Hilbert's involvement with both mathematics and physics.

On the mathematical side one may mention the intense research activity associated with the names of Cayley and Klein in projective geometry, concerning both the main body of results and the foundations of this discipline; the questions sparked by the discovery and publication of non-Euclidean geometries, which raised philosoph-

ical concerns to a larger extent than they elicited actual mathematical research; the introduction by Riemann of the manifold approach to the analysis of space and its elaboration by Lie and Helmholtz; the question of the arithmetization of the continuum as analyzed by Dedekind, which had also important foundational consequences for analysis; the gradual re-elaboration of axiomatic techniques and perspectives as a main approach to foundational questions in mathematics, especially in the hands of Grassmann and of the Italian geometers. Hilbert's intellectual debts to each of these traditions and to the mathematicians that partook in it – even though more complex and subtle than may appear on first sight – belong to the directly visible, received image of Hilbert the geometer.

What is remarkable, and virtually absent from the traditional historiography until relatively recently, is the extent to which similar parallel developments in physics played a fundamental role in shaping Hilbert's views on axiomatization. Very much like geometry, also physics underwent major changes throughout the nineteenth century. These changes affected the contents of the discipline, its methodology, its institutional setting, and its image in the eyes of its practitioners. They were accompanied by significant foundational debates that intensified considerably toward the end of the century, especially among German-speaking physicists. Part of these debates also translated into specific attempts to elucidate the role of basic laws or principles in physical theories, parallel in certain respects to that played by axioms in mathematical theories. As with geometry, foundational questions attracted relatively limited attention from practitioners of the discipline, but some leading figures were indeed involved in them.

From about 1850 on, physics became focused on quantification and the search for universal mathematical laws as its fundamental methodological principles, on the conservation of energy as a fundamental unifying principle, and very often on mechanical explanation of all physical phenomena as a preferred research direction. If explanations based on imponderable "fluids" had dominated so far, mechanical explanations based on the interaction of particles of ordinary matter now became much more frequent. In particular, the mechanical theory of ether gave additional impulse to the concept of "field" that would eventually require a mechanical explanation. Likewise, the kinetic theory of gases gave additional support to the foundational role of mechanics as a unifying, explanatory scheme. On the other hand, these very developments gave rise to many new questions that would eventually challenge the preferential status of mechanics and lead to the formulation of significant alternatives to it, especially in the form of the so-called "electromagnetic worldview", as well as in the "energicist" and the phenomenological approaches.

Beginning in the middle of the century, several physicists elaborated on the possibility of systematically clarifying foundational issues of this kind in physical theories, based on the use of "axioms", "postulates" or "principles". This was not, to be sure, a really central trend that engaged the leading physicists in lively discussions. Still, given the vivid interest on Volkmann in the topic, Hilbert became keenly aware of many of these developments and discussed them with his colleagues at the seminar.

Above all, the ideas of Heinrich Hertz and Ludwig Boltzmann on the foundations of physics strongly influenced him, not only at the methodological level, but also concerning his strong adherence to the mechanical reductionist point of view.

The lecture notes of courses in geometry taught by Hilbert in Königsberg illuminatingly exemplify the confluence of the various points mentioned in the preceding paragraphs. Central to this is his conception of geometry as a *natural* science, close in all respects to mechanics and the other physical disciplines, and opposed to arithmetic and other mathematical fields of enquiry. This was a traditional separation, adopted with varying degrees of commitment, among the German mathematicians (especially in Göttingen) since the time of Gauss. Even geometers like Moritz Pasch, who had stressed a thoroughly axiomatic approach in their presentations of projective geometry [20], would support such an empiricist view of geometry. In the introduction to a course taught in 1891, for instance, Hilbert expressed his views as follows:

> Geometry is the science dealing with the properties of space. It differs essentially from pure mathematical domains such as the theory of numbers, algebra, or the theory of functions. The results of the latter are obtained through pure thinking …The situation is completely different in the case of geometry. I can never penetrate the properties of space by pure reflection, much the same as I can never recognize the basic laws of mechanics, the law of gravitation or any other physical law in this way. Space is not a product of my reflections. Rather, it is given to me through the senses. ([5], p. 84)

The connection between this view and the axiomatic approach as a proper way to deal with this kind of sciences was strongly supported by the work of Hertz. Hilbert had announced another course in geometry for 1893, but for lack of students registered it was postponed until 1894. Precisely at this time, Hertz's *Principles of Mechanics* [13] was posthumous published, and Hilbert got enthusiastic notice of the book from his friend Minkowski. Minkowski had been in Bonn since 1885 where he came under the strong influence of Hertz, to the point that the latter became his main source of scientific inspiration ([15], p. 355). In the now famous introduction to his book, Hertz described physical theories as "pictures" (*Bilder*) that we form for ourselves of natural phenomena, and suggested three criteria to evaluate among several possible images of one and the same object: permissibility, correctness, and appropriateness. Permissibility corresponds very roughly to consistency, whereas correctness and appropriateness are closer to the kind of criteria that will appear later on in Hilbert's *Grundlagen der Geometrie* (*GdG* – see below).

In the lecture notes of his 1893–94 course, Hilbert referred once again to the natural character of geometry and explained the possible role of axioms in elucidating its foundations. As he had time to correct the notes, he now made explicit reference to Hertz's characterization of a "correct" scientific image (*Bild*) or theory. Thus Hilbert wrote ([5], p. 87):

> Nevertheless the origin [of geometrical knowledge] is in experience. The axioms are, as Hertz would say, images or symbols in our mind, such that consequents of the images are again images of the consequences, i.e., what we can logically deduce from the images is itself valid in nature.

Hilbert also pointed out the need of establishing the independence of the axioms of geometry, while alluding, once again, to the kind of demand stipulated by Hertz. Stressing the objective and factual character of geometry, Hilbert wrote:

> The problem can be formulated as follows: What are the necessary, sufficient, and mutually independent conditions that must be postulated for a system of things, in order that any of their properties correspond to a geometrical fact and, conversely, in order that a complete description and arrangement of all the geometrical facts be possible by means of this system of things.

The axioms of geometry and of physical disciplines, Hilbert said, "express observations of facts of experience, which are so simple that they need no additional confirmation by physicists in the laboratory."

The empirical character of geometry has its clear expression in the importance attributed to Gauss's measurement of the sum of angles of a triangle formed by three mountain peaks in Hannover. Hilbert found these measurements convincing enough to indicate the correctness of Euclidean geometry as a true description of physical space. Nevertheless, he envisaged the possibility that some future measurement would yield a different result. This example would arise very frequently in Hilbert's lectures on physics in years to come, as an example of how the axiomatic method should be applied in physics, where new empirical facts are often found by experiment. Faced with new such findings that seem to contradict an existing theory, the axiomatic analysis would allow making the necessary modifications on some of the basic assumptions of the theory, without however having to modify its essential logical structure. Hilbert stressed that the axiom of parallels is likely to be the one to be modified in geometry if new experimental discoveries would necessitate so. Geometry was especially amenable to a full axiomatic analysis only because of its very advanced stage of development and elaboration, and not because of any other specific, essential trait concerning its nature that would set it apart from other disciplines of physics. Thus, in a course on mechanics taught in 1899, the year of publication of *GdG*, he said:

> Geometry also [like mechanics] emerges from the observation of nature, from experience. To this extent, it is an *experimental science*.…But its experimental foundations are so irrefutably and so *generally acknowledged*, they have been confirmed to such a degree, that no further proof of them is deemed necessary. Moreover, all that is needed is to derive these foundations from a minimal set of *independent axioms* and thus to construct the whole edifice of geometry by *purely logical means*. In this way [i.e., by means of the axiomatic treatment] geometry is turned into a *pure mathematical* science. In mechanics it is also

> the case that all physicists recognize its most basic facts. But the *arrangement* of the basic concepts is still subject to changes in perception …and therefore mechanics cannot yet be described today as a *pure mathematical* discipline, at least to the same extent that geometry is. ([5], p. 90. Emphasis in the original)

Thus, at the turn of the century, Hilbert consolidated his view of the axiomatic method as a correct methodology to be applied, in parallel and with equal importance, to geometry and to all other physical disciplines. The publication of *GdG* helped spread his ideas very quickly and in strong association with geometry alone. But the idea of applying the same point of view to physics, although made known to the public only in the 1900 list of problems, was for him natural and evident from the outset. In his course of 1899, Hilbert devoted considerable effort to discussing the technical details of, as well as the logical and conceptual interrelations among, the main principles of analytical mechanics: the energy conservation principle, the principle of virtual velocities and the D'Alembert principle, the principles of straightest path and of minimal constraint, and the principles of Hamilton and Jacobi. All of this will appear prominently in Hilbert's later own elaboration of the program for the axiomatization of physics.

## 4. *Grundlagen der Geometrie*

Hilbert's *Grundlagen der Geometrie* embodied his first published, comprehensive presentation of an axiomatized mathematical discipline. Based on a course taught in the winter semester of 1898–99, it appeared in print in June of 1899. The declared aim of the book was to lay down a "simple" and "complete" system of "mutually independent" axioms, from which all known theorems of geometry might be deduced. The axioms were formulated for three systems of undefined objects named "points", "lines", and "planes", and they establish mutual relations that these objects must satisfy. The axioms were grouped into five categories: axioms of incidence, of order, of congruence, of parallels, and of continuity. From a purely logical point of view, the groups have no real significance in themselves. However, from the geometrical point of view they are highly significant, for they reflect Hilbert's actual conception of the axioms as an expression of spatial intuition: each group expresses a particular way that these intuitions manifest themselves in our understanding.

Hilbert's first requirement, that the axioms be independent, is the direct manifestation of the foundational concerns that guided his research. When analyzing independence, his interest focused mainly on the axioms of congruence, continuity and of parallels, since this independence would specifically explain how the various basic theorems of Euclidean and projective geometry are logically interrelated. This requirement had already appeared – albeit more vaguely formulated – in Hilbert's early lectures on geometry, as a direct echo of Hertz's demand for "appropriateness" of physical theories (i.e., the demand of "distinctness and simplicity" for the axioms

of the theory). This time Hilbert also provided the tools to prove systematically the mutual independence among the individual axioms within the groups and among the various groups of axioms in the system. However, this was not for Hilbert an exercise in analyzing abstract relations among systems of axioms and their possible models. The motivation for enquiring about the mutual independence of the axioms remained, essentially, a geometrical one. For this reason, Hilbert's original system of axioms was not the most economical one from the logical point of view. Indeed, several mathematicians noticed quite soon that Hilbert's system of axioms, seen as a single collection rather than as a collection of five groups, contained a certain degree of redundancy ([19], [23]). Hilbert's own aim was to establish the interrelations among the groups of axioms, embodying the various manifestations of space intuition, rather than among individual axioms belonging to different groups.

The second one, simplicity is also related to Hertz's appropriateness. Unlike the other requirements, it did not become standard as part of the important mathematical ideas to which *GdG* eventually led. Through this requirement Hilbert wanted to express the desideratum that an axiom should contain "no more than a single idea." However, he did not provide any formal criterion to decide when an axiom is simple. Rather this requirement remained implicitly present in *GdG*, as well as in later works of Hilbert, as a merely aesthetic guideline that was never transformed into a mathematically controllable feature.

The idea of a complete axiomatic system became pivotal to logic after 1930 following the works of Gödel, and in connection with the finitist program for the foundations of arithmetic launched by Hilbert and his collaborators around 1920. This is not, however, what Hilbert had in mind in 1899, when he included a requirement under this name in the analysis presented in *GdG*. Rather, he was thinking of a kind of "pragmatic" completeness. In fact, what Hilbert was demanding here is that an adequate axiomatization of a mathematical discipline should allow for an actual derivation of *all* the theorems already known in that discipline. This was, Hilbert claimed, what the totality of his system of axioms did for Euclidean geometry or, if the axiom of parallels is ignored, for the so-called absolute geometry, namely that which is valid independently of the latter.

Also the requirement of consistency was to become of paramount importance thereafter. Still, as part of *GdG*, Hilbert devoted much less attention to it. For one thing, he did not even mention this task explicitly in the introduction to the book. For another, he devoted just two pages to discussing the consistency of his system in the body of the book. In fact, it is clear that Hilbert did not intend to give a direct proof of consistency of geometry here, but even an indirect proof of this fact does not explicitly appear in *GdG*, since a systematic treatment of the question implied a full discussion of the structure of the system of real numbers, which was not included. Rather, Hilbert suggested that it would suffice to show that the specific kind of synthetic geometry derivable from his axioms could be translated into the standard Cartesian geometry, if the axes are taken as representing the entire field of real numbers. Only in the second edition of *GdG*, published in 1903, Hilbert added an additional axiom,

the so-called "axiom of completeness" (*Vollständigkeitsaxiom*), meant to ensure that, although infinitely many incomplete models satisfy all the other axioms, there is only one complete model that satisfies this last axiom as well, namely, the usual Cartesian geometry.

Hilbert's axiomatic analysis of geometry was not meant to encourage the possibility of choosing arbitrary combinations of axioms within his system, and of exploring their consequences. Rather, his analysis was meant to enhance our understanding of those systems with a more intuitive, purely geometrical significance – Euclidean geometry, above all – and that made evident the connection of his work with long-standing concerns of the discipline throughout the nineteenth century [8]. As already stressed, the definition of systems of abstract axioms and the kind of axiomatic analysis described above was meant to be carried out always retrospectively, and only for "concrete", *well-established and elaborated* mathematical entities.

The publication of the *Grundlagen* was followed by many further investigations into Hilbert's technical arguments, as well as by more general, methodological and philosophical discussions. One important such discussion appeared in the correspondence between Hilbert and Gottlob Frege. This interchange has drawn considerable attention of historians and philosophers, especially for the debate it contains between Hilbert and Frege concerning the nature of mathematical truth. But this frequently-emphasized issue is only one side of a more complex picture advanced by Hilbert in his letters. In particular, it is interesting to notice Hilbert's explanation to Frege, concerning the main motivations for undertaking his axiomatic analysis: the latter had arisen, in the first place, from difficulties Hilbert had encountered when dealing with *physical*, rather than mathematical theories. Echoing once again ideas found in the introduction to Hertz's textbook, and clearly having in mind the problematic conceptual situation of the kinetic theory of gases at the turn of the century, Hilbert stressed the need to analyze carefully the process whereby physicists continually add new assumptions to existing physical theories, without properly checking whether or not the former contradict the latter, or consequences of the latter. In a letter of December 29, 1899, Hilbert wrote to Frege:

> After a concept has been fixed completely and unequivocally, it is on my view completely illicit and illogical to add an axiom – a mistake made very frequently, especially by physicists. By setting up one new axiom after another in the course of their investigations, without confronting them with the assumptions they made earlier, and without showing that they do not contradict a fact that follows from the axioms they set up earlier, physicists often allow sheer nonsense to appear in their investigations. One of the main sources of mistakes and misunderstandings in modern physical investigations is precisely the procedure of setting up an axiom, appealing to its truth, and inferring from this that it is compatible with the defined concepts. One of the main purposes of my *Festschrift* was to avoid this mistake. ([9], p. 40)

In a different passage of the same letter, Hilbert commented on the possibility of substituting the basic objects of an axiomatically formulated theory by a different system of objects, provided the latter can be put in a one-to-one, invertible relation with the former. In this case, the known theorems of the theory are equally valid for the second system of objects. Concerning physical theories, Hilbert wrote:

> All the statements of the theory of electricity are of course valid for any other system of things which is substituted for the concepts magnetism, electricity, etc., provided only that the requisite axioms are satisfied. But the circumstance I mentioned can never be a defect in a theory [footnote: it is rather a tremendous advantage], and it is in any case unavoidable. However, to my mind, the application of a theory to the world of appearances always requires a certain measure of good will and tactfulness: e.g., that we substitute the smallest possible bodies for points and the longest possible ones, e.g., light-rays, for lines. At the same time, the further a theory has been developed and the more finely articulated its structure, the more obvious the kind of application it has to the world of appearances, and it takes a very large amount of ill will to want to apply the more subtle propositions of [the theory of surfaces] or of Maxwell's theory of electricity to other appearances than the ones for which they were meant …([9], p. 41)

Hilbert's letters to Frege help understanding the importance of the link between physical and mathematical theories on the development of his axiomatic point of view. The latter clearly did not involve either an empty game with arbitrary systems of postulates nor a conceptual break with the classical, nineteenth-century entities and problems of mathematics and empirical science. Rather it sought after an improvement in the mathematician's understanding of the latter. This motto was to guide much of Hilbert's incursions into several domains of physics over the years to come.

## 5. Physics and the 1900 list of problems

In the introductory section of his Paris talk, Hilbert stressed the important role he accorded to empirical motivations as a fundamental source of nourishment for what he described as a "living organism", in which mathematics and the physical sciences appear tightly interrelated. The empirical motivations underlying mathematical ideas, Hilbert said, should by no means be taken as opposed to rigor. On the contrary, contrasting an "opinion occasionally advocated by eminent men", Hilbert insisted that the contemporary quest for rigor in analysis and arithmetic should in fact be *extended to both geometry and the physical sciences*. He was alluding here, most probably, to Kronecker and Weierstrass, and the Berlin purist tendencies that kept geometry and applications out of their scope of interest. Rigorous methods are often simpler and easier to understand, Hilbert said, and therefore, a more rigorous treatment would

only perfect our understanding of these topics, and at the same time would provide mathematics with ever new and fruitful ideas. In explaining why rigor should not be sought only within analysis, Hilbert actually implied that this rigor should actually be pursued in axiomatic terms. He thus wrote:

> Such a one-sided interpretation of the requirement of rigor would soon lead to the ignoring of all concepts arising from geometry, mechanics and physics, to a stoppage of the flow of new material from the outside world, and finally, indeed, as a last consequence, to the rejection of the ideas of the continuum and of irrational numbers. But what an important nerve, vital to mathematical science, would be cut by rooting out geometry and mathematical physics! On the contrary I think that wherever mathematical ideas come up, whether from the side of the theory of knowledge or in geometry, or from the theories of natural or physical science, the problem arises for mathematics to investigate the principles underlying these ideas and to establish them upon a simple and complete system of axioms, so that the exactness of the new ideas and their applicability to deduction shall be in no respect inferior to those of the old arithmetical concepts. (Quoted from [12], p. 245)

Using a rhetoric reminiscent of Volkmann's work, Hilbert described the development of mathematical ideas as an ongoing, dialectical interplay between the two poles of thought and experience. He also added an idea that was of central importance to Göttingen scientists for many decades, namely, the conception of the "pre-established harmony" between mathematics and nature ([21]). The importance of investigating the foundations of mathematics does not appear as an isolated concern, but rather as an organic part of the manifold growth of the discipline in several directions. Hilbert thus said:

> Indeed, the study of the foundations of a science is always particularly attractive, and the testing of these foundations will always be among the foremost problems of the investigator …[But] a thorough understanding of its special theories is necessary for the successful treatment of the foundations of the science. Only that architect is in the position to lay a sure foundation for a structure who knows its purpose thoroughly and in detail. (Quoted from [12], p. 258)

The first two problems in Hilbert's list are Cantor's continuum hypothesis and the compatibility of the axioms of arithmetic. In formulating the second problem on his list, Hilbert stated more explicitly than ever before, that among the tasks related to investigating an axiomatic system, proving its consistency would be the most important one. Yet, Hilbert was still confident that this would be a rather straightforward task, easily achievable "by means of a careful study and suitable modification of the known methods of reasoning in the theory of irrational numbers." Clearly Hilbert meant his remarks in this regard to serve as an argument against Kronecker's negative

reactions to unrestricted use of infinite collections in mathematics, and therefore he explicitly asserted that a consistent system of axioms could prove the existence of higher Cantorian cardinals and ordinals. Hilbert's assertion is actually the first published mention of the paradoxes of Cantorian set theory, which here were put forward with no special fanfare ([7], p. 301). He thus established a clear connection between the two first problems on his list through the axiomatic approach. Still, Hilbert was evidently unaware of the difficulties involved in realizing this point of view, and, more generally, he most likely had no precise idea of what an elaborate theory of systems of axioms would involve. On reading the first draft of the Paris talk, several weeks earlier, Minkowski understood at once the challenging implications of Hilbert's view, and he hastened to write to his friend:

> In any case, it is highly original to proclaim as a problem for the future, one that mathematicians would think they had already completely possessed for a long time, such as the axioms for arithmetic. What might the many laymen in the auditorium say? Will their respect for us grow? And you will also have a though fight on your hands with the philosophers. ([22], p. 129)

Frege's reaction to the *GdG* proved Minkowski's concern to be justified, as his main criticism referred to the status of axioms as implicit definitions.

The next three problems in the list are directly related with geometry and, although not explicitly formulated in axiomatic terms, they address the question of finding the correct relationship between specific assumptions and specific, significant geometrical facts. The fifth problem, for instance, relates to the question of the foundations of geometry as it had evolved over the last third of the nineteenth century along two parallel paths. On the one hand, there was the age-old tradition of elementary synthetic geometry, where the question of foundations more naturally arises in axiomatic terms. On the other hand, there was the tradition associated with the Helmholtz–Lie problem, that derived directly from the work of Riemann and that had a more physically-grounded orientation connected with the question of spaces that admit the free mobility of rigid bodies. Whereas Helmholtz had only assumed *continuity* as underlying the motion of rigid bodies, in applying his theory of groups of transformations to this problem, Lie was also assuming the *differentiability* of the functions involved. Hilbert's work on the foundations of geometry, especially in the context that led to *GdG*, had so far been connected with the first of these two traditions, while devoting much less attention to the second one. Now in his fifth problem, he asked whether Lie's conditions, rather than assumed, could actually be deduced from the group concept together with other geometrical axioms.

As a mathematical problem, the fifth one led to interesting, subsequent developments. Not long after his talk, in November 18, 1901, Hilbert himself proved that, in the plane, the answer is positive, and he did so with the help of a then innovative, essentially topological, approach [14]. That the answer is positive in the general case was satisfactorily proved only in 1952 ([10], [18]). The inclusion of this problem in

the 1900 list underscores the actual scope of Hilbert's views over the question of the foundations of geometry and over the role of axiomatics. Hilbert suggested here the pursuit of an intricate kind of conceptual clarification involving assumptions about motion, differentiability and symmetry, such as they appear intimately interrelated in the framework of a well-elaborate mathematical theory, namely, that of Lie. This quest, that also became typical of the spirit of Hilbert's axiomatic involvement with physical theories, suggests that his foundational views on geometry were very broad and open-ended, and did not focus on those aspects related with the synthetic approach to geometry. In particular, the fifth problem emphasizes the prominent role that Hilbert assigned to physical considerations in his approach to geometry. In the long run, this aspect of Hilbert's view resurfaced at the time of his involvement with GTR ([5], Ch. 7–8). In its more immediate context, however, it makes the passage from geometry to the sixth problem appear as a natural one within the list.

Indeed, if the first two problems in the list show how the ideas deployed in *GdG* led in one direction towards foundational questions in arithmetic, then the fifth problem suggests how they also naturally led, in a different direction, to Hilbert's call for the axiomatization of physical science in the sixth problem. The problem was thus formulated as follows:

> The investigations on the foundations of geometry suggest the problem: To treat in the same manner, by means of axioms, those physical sciences in which mathematics plays an important part; in the first rank are the theory of probabilities and mechanics. (Quoted in [12], p. 258)

As examples of what he had in mind Hilbert mentioned several existing and well-known works: the fourth edition of Mach's *Die Mechanik in ihrer Entwicklung*, Hertz's *Prinzipien*, Boltzmann's 1897 *Vorlesungen Über die Principien der Mechanik*, and also Volkmann's 1900 *Einführung in das Studium der theoretischen Physik*. Boltzmann's work offered a good example of what axiomatization would offer, as he had indicated, though only schematically, that limiting processes could be applied, starting from an atomistic model, to obtain the laws of motion of continua. Hilbert thought it convenient to go in the opposite direction also, i.e., to derive the laws of motions of rigid bodies by limiting processes, starting from a system of axioms that describe space as filled with continuous matter in varying conditions. Thus one could investigate the equivalence of different systems of axioms, an investigation that Hilbert considered to be of the highest theoretical importance.

This is one of the few places where Hilbert emphasized Boltzmann's work over Hertz's in this regard, and this may give us the clue to the most immediate trigger that was in the back of Hilbert's mind when he decided to include this problem in the list. Indeed, Hilbert had met Boltzmann several months earlier in Munich, where the latter gave a talk on recent developments in physics. Boltzmann had not only discussed ideas connected with the task that Hilbert was now calling for, but he also adopted a rhetoric that seems to have appealed very much to Hilbert. In fact, Boltzmann

had suggested that one could follow up the recent history of physics with a look at future developments. Nevertheless, he said, "I will not be so rash as to lift the veil that conceals the future" ([2], p. 79). Hilbert, on the contrary, opened the lecture by asking precisely, "who among us would not be glad to lift the veil behind which the future lies hidden" and the whole trust of his talk implied that he, the optimistic Hilbert, was helping the mathematical community to do so.

Together with the well-known works on mechanics referred to above, Hilbert also mentioned a recent work by the Göttingen actuarial mathematician Georg Bohlmann on the foundations of the calculus of probabilities [1]. The latter was important for physics, Hilbert said, for its application to the method of mean values and to the kinetic theory of gases. Hilbert's inclusion of the theory of probabilities among the main physical theories whose axiomatization should be pursued has often puzzled readers of this passage. The notes of a course taught in 1905 on the axiomatic method show that this was a main point in Hilbert's views on physics because of the use of probabilities also in insurance mathematics and in problems of observational error calculation in astronomy. It is also remarkable that Hilbert did not mention electrodynamics among the physical disciplines to be axiomatized, even though the second half of the *Gauss–Weber Festschrift*, where Hilbert's *GdG* was published, contained a parallel essay by Wiechert on the foundations of electrodynamics. At any rate, Wiechert's presentation was by no means axiomatic, in any sense of the term. On the other hand, the topics addressed by Wiechert would start attracting Hilbert's attention over the next years, at least since 1905.

This sixth problem is not really a problem in the strict sense of the word, but rather a general task for whose complete fulfillment Hilbert set no clear criteria. Thus, Hilbert's detailed account in the opening remarks of his talk as to what a meaningful problem in mathematics is, and his stress on the fact that a solution to a problem should be attained in a finite number of steps, does not apply in any sense to the sixth one. On the other hand, the sixth problem has important connections with three other problems on Hilbert's list: the nineteenth ("Are all the solutions of the Lagrangian equations that arise in the context of certain typical variational problems necessarily analytic?"), the twentieth (dealing with the existence of solutions to partial differential equations with given boundary conditions), closely related to the nineteenth and at the same time to Hilbert's long-standing interest on the Dirichlet Principle, and, finally, the twenty-third (an appeal to extend and refine the existing methods of variational calculus). Like the sixth problem, the latter two are general tasks rather than specific mathematical problems with a clearly identifiable, possible solution. All these three problems are also strongly connected to physics, though unlike the sixth, they are also part of mainstream, traditional research concerns in mathematics. In fact, their connections to Hilbert's own interests are much more perspicuous and, in this respect, they do not raise the same kind of historical questions that Hilbert's interest in the axiomatization of physics does.

A balanced assessment of the influence of the problems on the development of mathematics throughout the century must take into account not only the intrinsic

importance of the problems, but also the privileged institutional role of Göttingen in the mathematical world with the direct and indirect implications of its special status. However, if Hilbert wished to influence the course of mathematics over the coming century with his list, then it is remarkable that his own career was only very partially shaped by it. Part of the topics covered by the list belonged to his previous domains of research, while others belonged to domains where he never became active. On the contrary, domains that he devoted much effort to over the next years, such as the theory of integral equations, were not contemplated in the list. In spite of the enormous influence Hilbert had on his students, the list did not become a necessary point of reference of preferred topics for dissertations. To be sure, some young mathematicians, both in Göttingen and around the world, did address problems on the list and sometimes came up with important mathematical achievements that helped launch their own international careers. But this was far from the only way for talented young mathematicians to reach prominence in or around Göttingen. But, ironically, the sixth problem, although seldom counted among the most influential of the list, can actually be counted among those that received greater attention from Hilbert himself and from his collaborators and students over the following years.

## 6. Concluding remarks

For all its differences and similarities with other problems on the list, the important point that emerges from the above account is that the sixth problem was in no sense disconnected from the evolution of Hilbert's early axiomatic conception at its very core. Nor was it artificially added in 1900 as an afterthought about the possible extensions of an idea successfully applied in 1899 to the case of geometry. Rather, Hilbert's ideas concerning the axiomatization of physical science arose simultaneously with his increasing enthusiasm for the axiomatic method and they fitted naturally into his overall view of pure mathematics, geometry and physical science – and the relationship among them – by that time.

From 1900 on, the idea of axiomatizing physical theories was a main thread that linked much of Hilbert's research and teaching. Hilbert taught every semester at least one course dealing with a physical discipline, and by the end of his career he had covered most of the important fields that were at the cutting edge of physics, currently attracting the best research efforts of young and promising minds (see the appendix to this article). The axiomatic point of view provided a unifying methodology from which to approach many of the topics in which Hilbert became interested. In 1905 he taught a course on the axiomatic method where he presented for the first time a panoramic view of various physical disciplines from an axiomatic perspective: mechanics, thermodynamics, probability calculus, kinetic theory, insurance mathematics, electrodynamics, psychophysics. The variety of physical topics pursued only grew over the years. The extent of the influence of Hilbert's ideas on physics on contemporary research is a more complex question that cannot be discussed here for lack

of space. Still, it is relevant to quote here an account of Hilbert's ideas as described by the physicist on whom Hilbert's influence became most evident, Max Born. On the occasion of Hilbert's sixtieth birthday, at a time when he was deeply involved together with Bernays in the technical difficulties raised by the finitist program, Born wrote the following words:

> The physicist set outs to explore how things are in nature; experiment and theory are thus for him only a means to attain an aim. Conscious of the infinite complexities of the phenomena with which he is confronted in every experiment, he resists the idea of considering a theory as something definitive. He therefore abhors the word "Axiom", which in its usual usage evokes the idea of definitive truth. The physicist is thus acting in accordance with his healthy instinct, that dogmatism is the worst enemy of natural science. The mathematician, on the contrary, has no business with factual phenomena, but rather with logic interrelations. In Hilbert's language the axiomatic treatment of a discipline implies in no sense a definitive formulation of specific axioms as eternal truths, but rather the following methodological demand: specify the assumptions at the beginning of your deliberation, stop for a moment and investigate whether or not these assumptions are partly superfluous or contradict each other. ([3])

The development of physics from the beginning of the century, and especially *after* 1905, brought many surprises that Hilbert could not have envisaged in 1900 or even when he lectured at Göttingen on the axioms of physics in 1905; yet, Hilbert was indeed able to accommodate these new developments to the larger picture of physics afforded by his program for axiomatization. In fact, some of his later contributions to mathematical physics, particularly his contributions to GTR, came by way of realizing the vision embodied in this program.

## 7. Appendix: Hilbert's Göttingen courses on physics (and related fields): 1895–1927

For an explanation on the sources used for compiling this list, see [5], p. 450 (WS = Winter Semester, SS = Summer Semester, HS = Special Autumn [Herbst] Semester).

| | |
|---|---|
| **WS 1895/96** | Partial Differential Equations |
| **SS 1896** | Ordinary Differential Equations |
| **SS 1898** | Mechanics |
| **SS 1899** | Variational Calculus |
| **WS 1900/01** | Partial Differential Equations |
| **SS 1901** | Linear Partial Differential Equations |
| **WS 1901/02** | Potential Theory |

| | |
|---|---|
| **SS 1902** | Selected Topics in Potential Theory |
| **WS 1902/03** | Continuum Mechanics - Part I |
| **SS 1903** | Continuum Mechanics - Part II |
| **WS 1903/04** | Partial Differential Equations |
| **WS 1904/05** | Variational Calculus |
| **SS 1905** | Logical Principles of Mathematical Thinking (and of Physics) |
| **SS 1905** | Integral Equations |
| **WS 1905/06** | Partial Differential Equations |
| **WS 1905/06** | Mechanics |
| **SS 1906** | Integral Equations |
| **WS 1906/07** | Continuum Mechanics |
| **SS 1907** | Differential Equations |
| **WS 1909/10** | Partial Differential Equations |
| **SS 1910** | Selected Chapters in the Theory of Partial Differential Equations |
| **WS 1910/11** | Mechanics |
| **SS 1911** | Continuum Mechanics |
| **WS 1911/12** | Statistical Mechanics |
| **SS 1912** | Radiation Theory |
| **SS 1912** | Ordinary Differential Equations |
| **SS 1912** | Mathematical Foundations of Physics |
| **WS 1912/13** | Molecular Theory of Matter |
| **WS 1912/13** | Partial Differential Equations |
| **WS 1912/13** | Mathematical Foundations of Physics |
| **SS 1913** | Foundations of Mathematics (and the axiomatization of Physics) |
| **SS 1913** | Electron Theory |
| **WS 1913/14** | Electromagnetic Oscillations |
| **WS 1913/14** | Analytical Mechanics |
| **WS 1913/14** | Exercises in Mechanics (together with H. Weyl) |
| **SS 1914** | Statistical Mechanics |
| **SS 1914** | Differential Equations |
| **WS 1914/15** | Lectures on the Structure of Matter |
| **SS 1915** | Structure of Matter (Born's Theory of Crystals) |
| **WS 1915/16** | Differential Equations |
| **SS 1916** | Partial Differential Equations |
| **SS 1916** | Foundations of Physics I (General Relativity) |
| **WS 1916/17** | Foundations of Physics II (General Relativity) |
| **SS 1917** | Electron Theory |
| **SS 1918** | Ordinary Differential Equations |
| **WS 1918/19** | Space and Time |
| **WS 1918/19** | Partial Differential and Integral Equations |
| **HS 1919** | Nature and Mathematical Knowledge |
| **WS 1920** | Mechanics |
| **SS 1920** | Higher Mechanics and the New Theory of Gravitation |

| | |
|---|---|
| **WS 1920/21** | Mechanics and the New Theory of Gravitation |
| **SS 1921** | Einstein's Gravitation Theory. Basic Principles of the Theory of Relativity |
| **SS 1921** | On Geometry and Physics |
| **SS 1922** | Statistical Mechanics |
| **WS 1922/23** | Mathematical Foundations of Quantum Theory |
| **WS 1922/23** | Knowledge and Mathematical Thought |
| **WS 1922/23** | Knowledge and Mathematical Thought |
| **SS 1923** | Our Conception of Gravitation and Electricity |
| **WS 1923/24** | On the Unity of Science |
| **SS 1924** | Mechanics and Relativity Theory |
| **WS 1926/27** | Mathematical Methods of Quantum Theory |
| **SS 1930** | Mathematical Methods of Modern Physics |
| **WS 1930/31** | Nature and Thought |
| **WS 1931/32** | Philosophical Foundations of Modern Natural Science |

## References


[1] Bohlmann, G., Ueber Versicherungsmathematik. In *Über angewandte Mathematik und Physik in ihrer Bedeutung für den Unterricht an den höheren Schulen* (ed. by F. Klein & E. Riecke), Teubner, Leipniz, Berlin 1900, 114–145.

[2] Boltzmann, L., Über die Entwicklung der Methoden der theoretischen Physik in neuerer Zeit (1899). In L. Boltzmann *Populäre Schriften*, J. A. Barth, Leipzig 1905, 198–277.

[3] Born, M., Hilbert und die Physik. *Die Naturwissenschaften* 10 (1922), 88–93. (Reprint in Born. M., *Ausgewählte Abhandlungen*, Vol. 2, Vandenhoek & Ruprecht, Göttingen 1963, 584–598.)

[4] Brush, S. G., *The Kind of Motion we Call Heat - A History of the Kinetic Theory of Gases in the 19th Century*. North Holland Publishing House, Amsterdam, New York, Oxford 1976.

[5] Corry, L., *David Hilbert and the Axiomatization of Physics (1898–1918)*: From *Grundlagen der Geometrie* to *Grundlagen der Physik.* Archimedes: New Studies in the History and Philosophy of Science and Technology 10, Kluwer Academic Publishers, Dordrecht 2004.

[6] Dieudonné, J., Les méthodes axiomatiques modernes et les fondements des mathématiques. In *Les grands Courants de la Pensée Mathématique* (ed. by F. Le Lionnais), Blanchard, Paris 1962, 443–555.

[7] Ferreirós, J., *Labyrinths of Thought. A History of Set Theory and its Role in Modern Mathematics*. Sci. Networks Hist. Stud. 23, Birkhäuser, Boston 1999.

[8] Freudenthal, H., Zur Geschichte der Grundlagen der Geometrie. Zugleich eine Besprechung der 8. Auflage von Hilberts 'Grundlagen der Geometrie'. *Nieuw Archief voor Wiskunde* **4** (1957), 105–142.

[9] Gabriel, G. et al. (eds.), *Gottlob Frege - Philosophical and Mathematical Correspondence*. The University of Chicago Press, Chicago 1980.

[10] Gleason, A., Groups without Small Subgroups. *Ann. Math.* **56** (1952), 193–212.

[11] Gnedenko, J., Zum sechsten Hilbertschen Problem. In *Die Hilbertsche Probleme* (ed. by P. Alexandrov), Ostwalds Klassiker der exakten Wissenschaften 252, Leipzig 1979, 144–147.

[12] Gray, J. J., *The Hilbert Challenge*. Oxford University Press, New York 2000.

[13] Hertz, H., *Die Prinzipien der Mechanik in neuem Zusammenhange dargestellt*. Leipzig 1984.

[14] Hilbert, D., Über die Grundlagen der Geometrie. *Math. Ann.* **56** (1902), 233–241.

[15] Hilbert, D., Hermann Minkowski. *Math. Ann.* **68** (1910), 445–471.

[16] Hilbert, D., *Natur und Mathematisches Erkennen: Vorlesungen, gehalten 1919-1920 in Göttingen. Nach der Ausarbeitung von Paul Bernays*. Edited and with an English introduction by David E. Rowe, Birkhäuser, Basel 1992.

[17] Hilbert, D., Über meine Tätigkeit in Göttingen. In *Hilbert: Gedenkband* (ed. by K. Reidemeister). Springer-Verlag, Berlin, Heidelberg, New York 1971, 79–82.

[18] Montgomery, D., Zippin, L., Small Subgroups of Finite-dimensional Groups. *Ann. Math.* **56** (1952), 213–241.

[19] Moore. E. H., Projective Axioms of Geometry. *Trans. Amer. Math. Soc.* **3** (1902), 142–158.

[20] Pasch, M., *Vorlesungen über neuere Geometrie*. Teubner, Leipzig 1882.

[21] Pyenson, L., Relativity in Late Wilhelmian Germany: the Appeal to a Pre-established Harmony Between Mathematics and Physics. In L. Pyenson *The Young Einstein: The Advent of Relativity*, Adam Hilger Ltd., Bristol, Boston 1985, 137–157.

[22] Rüdenberg L., Zassenhaus, H., *Hermann Minkowski - Briefe an David Hilbert*. Springer-Verlag, Berlin, New York 1973.

[23] Schur, F., Über die Grundlagen der Geometrie. *Math. Ann.* **55** (1901), 265–292.

[24] Weyl, H., David Hilbert and his Mathematical Work, *Bull. Amer. Math. Soc.* **50** (1944), 612–654.

[25] Wightman, A. S., Hilbert's Sixth Problem: Mathematical Treatment of the Axioms of Physics. In *Mathematical Developments Arising from Hilbert Problems* (ed. by F. E. Browder), Symposia in Pure Mathematics 28, Amer. Math. Soc., Providence, RI, 1976, 147–240.

[26] Yandell, B. H., *The Honors Class: Hilbert's Problems and Their Solvers*. AK Peters, Natick, MA, 2002.

Cohn Institute for History and Philosophy of Science, Tel-Aviv University, Tel-Aviv 69978, Israel

E-mail: corry@post.tau.ac.il

# Method versus calculus in Newton's criticisms of Descartes and Leibniz

Niccolò Guicciardini

**Abstract.** In my talk I will consider Newton's views on mathematical method. Newton never wrote extensively about this issue. However, in his polemic writings addressed against Descartes and Leibniz he expressed the idea that his method was superior to the ones proposed by the French and the German. Considering these writings can help us in understanding the role attributed to algebra and calculus in Newton's mathematical thought.

**Mathematics Subject Classification (2000).** Primary 01A45; Secondary 00A30.

**Keywords.** 17th century, philosophy of mathematics, Newton, Descartes, Leibniz.

## 1. Newton's memorandum on his early discoveries

Newton blossomed as a creative mathematician in 1665–1666, the so-called *anni mirabiles*, about four years after matriculating at Cambridge.[1] A Newtonian memorandum, written about fifty years later, gives an account that has been basically confirmed by manuscript evidence:

> In the beginning of the year 1665 I found the Method of approximating series & the Rule for reducing any dignity of any Binomial into such a series. The same year in May I found the method of Tangents of Gregory & Slusius, & in November had the direct method of fluxions & the next year in January had the theory of Colours & in May following I had entrance into y$^{e}$ inverse method of fluxions. And the same year I began to think of gravity extending to y$^{e}$ orb of the Moon [...] All this was in the two plague years of 1665–1666. For in those days I was in the prime of my age for invention & minded Mathematicks & Philosophy more than any time since. ([1])

There would be much to say in order to decipher and place into context Newton's discourse. For instance, the task of commenting on the meaning of the term 'philosophy' would require space and scholarship not at my disposal [2].

[1] Readers interested in Newton's mathematics should read Tom Whiteside's introductions and commentaries in [9].

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006


Figure 1. Newton's home at Woolsthorpe where – he claimed – he made his early discoveries in mathematics and natural philosophy when Cambridge University was evacuated because of the plague during the biennium 1665–1666. As a matter of fact, he did important work in mathematics during periods in which he returned to the University. Further, his juvenile insights – particularly those concerning gravitation – had to be elaborated during the next decades. Source: [1], 54.

Let me note three things about the above memorandum. The 'Method of approximating series' is the method of series expansion via long division and root extraction (as well as other methods which were later subsumed under more general techniques usually attributed to Puiseux) that allowed Newton to go beyond the limitation of what he termed 'common analysis' – where 'finite equations' were deployed – and express certain curves locally in terms of infinite fractional power series, which Newton called 'infinite equations'. The 'Rule for reducing any dignity of any Binomial' is what we call the 'binomial theorem'. Such methods of series expansion were crucial for attaining two goals: the calculation of areas of curvilinear surfaces and the rectification of curves (see Figure 2). Notice that Newton does not talk about a theorem, but rather about 'methods' and a 'rule'. This last fact is of utmost importance and deserves our commentary in Sections 2, 3, and 4, before turning in Section 5 to the direct and inverse methods of fluxions which are the Newtonian equivalent of the Leibnizian differential and integral calculus.[2]

[2]For a recent evaluation of Newton's early mathematical researches see [3].

*Examples, where the Square Root muſt be extracted.*

15. If it be $\sqrt{aa+xx}=y$, I extract the Root thus:

$aa+xx\left(a+\frac{x^2}{2a}-\frac{x^4}{8a^3}+\frac{x^6}{16a^5}-\frac{5x^8}{128a^7}\ \&c.\right.$

$aa$

$0+x^2$

$x^2+\frac{x^4}{4a^2}$

$0-\frac{x^4}{4a^2}$

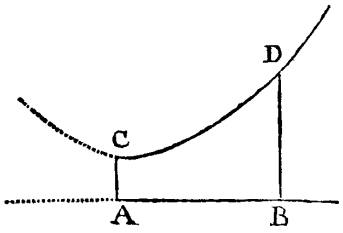
$-\frac{x^4}{4a^2}-\frac{x^6}{8a^4}+\frac{x^8}{64a^6}$

$0+\frac{x^6}{8a^4}-\frac{x^8}{64a^6}$

$+\frac{x^6}{8a^4}+\frac{x^8}{16a^6}-\frac{x^{10}}{64a^8}+\frac{x^{12}}{256a^{10}}$

$0-\frac{5x^8}{64a^6}+\frac{x^{10}}{64a^8}-\frac{x^{12}}{256a^{10}}$

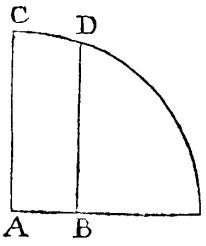
$\&c.$




Whence for the Equation $\sqrt{aa+xx}=y$, a new one is produced, *viz.* $y=a+\frac{x^2}{2a}-\frac{x^4}{8a^3}+\frac{x^6}{16a^5}-\frac{5x^8}{128a^7}\ \&c.$ And (by the ſecond Rule) You will have the Area ſought ABDC $=ax+\frac{x^3}{6a}-\frac{x^5}{40a^3}+\frac{x^7}{112a^5}-\frac{5x^9}{1152a^7}\ \&c.$

And this is the Quadrature of the Hyperbola.

16. After the ſame Manner if it be $\sqrt{aa-xx}=y$, it's Root will be $a-\frac{x^2}{2a}-\frac{x^4}{8a^3}-\frac{x6}{16a^5}-\frac{5x^8}{128a^7}\ \&c.$ and therefore the Area ſought ABDC will be equal to $ax-\frac{x^3}{6a}-\frac{x^5}{40a^3}-\frac{x^7}{112a^5}-\frac{5x^9}{1152a^7}\ \&c.$ And this is the Quadrature of the Circle.




17.

Figure 2. Calculation of areas of hyperbolic and circular surfaces via extraction of root of $\sqrt{aa+xx}=y$ and $\sqrt{aa-xx}=y$. This technique of series expansion and termwise integration was basic in Newton's early mathematical work and was displayed in a tract entitled *On the analysis by means of infinite equations* (written in 1669, but printed only in 1711), an extension of 'common analysis' which proceeds via 'finite equations' only. Source: [8], vol. 1, 8.

## 2. Pappus on the method of analysis and synthesis

Newton belonged to a mathematical community in which the distinction between theorems and problems was articulated according to criteria sanctioned by the venerated Greek tradition. Most notably in the work of the late Hellenistic compiler Pappus entitled *Mathematical Collection* which appeared in 1588 in Latin translation Newton – who avidly read this dusty work – could find a distinction between 'theorematic and problematic analysis'.

In the 7th book of the *Collection* there was a description of works (mostly lost and no longer available to early modern mathematicians) which – according to Pappus – had to do with a heuristic method followed by the ancient geometers. The opening of the seventh book is often quoted. It is an obscure passage whose decoding was top in the agenda of early modern European mathematicians, convinced as they were that here lay hidden the key to the method of discovery of the ancients. Given the importance this passage had for Newton, it is worth quoting at length:

> That which is called the *Domain of Analysis*, my son Hermodorus, is, taken as a whole, a special resource that was prepared, after the composition of the *Common Elements*, for those who want to acquire a power in geometry that is capable of solving problems set to them; and it is useful for this alone. It was written by three men: Euclid the Elementarist, Apollonius of Perge, and Aristaeus the elder, and its approach is by analysis and synthesis.
>
> Now analysis is the path from what one is seeking, as if it were established, by way of its consequences, to something that is established by synthesis. That is to say, in analysis we assume what is sought as if it has been achieved, and look for the thing from which it follows, and again what come before that, until by regressing in this way we come upon some one of the things that are already known, or that occupy the rank of a first principle. We call this kind of method 'analysis', as if to say *anapalin lysis* (reduction backward). In synthesis, by reversal, we assume what was obtained last in the analysis to have been achieved already, and, setting now in natural order, as precedents, what before were following, and fitting them to each other, we attain the end of the construction of what was sought.
>
> There are two kinds of analysis: one of them seeks after the truth, and is called 'theorematic': while the other tries to find what was demanded, and is called 'problematic'. In the case of the theorematic kind, we assume what is sought as a fact and true, then advancing through its consequences, as if they are true facts according to the hypothesis, to something established, if this thing that has been established is a truth, then that which was sought will also be true, and its proof the reverse of the analysis; but if we should meet with something established to be false, then the thing that was sought too will be false. In the case of the problematic kind, we assume the proposition as something we know, then, proceeding through its consequences, as if true, to something

> established, if the established thing is possible and obtainable, which is what mathematicians call 'given', the required thing will also be possible, and again the proof will be the reverse of the analysis; but should we meet with something established to be impossible, then the problem too will be impossible. ([4])

Pappus here made a distinction between analysis and synthesis. Analysis ('resolutio' in Latin) was often conceived of as a method of discovery, or a method of problem solving, which, working step by step backwards from what is sought as if it had already been achieved, eventually arrives at what is known. Synthesis ('compositio' or 'constructio') goes the other way round: it starts from what is known and, working through the consequences, arrives at what is sought. On the basis of Pappus' authority it was often stated that synthesis 'reverses' the steps of analysis. It was synthesis which provided the rigorous proof. Thus the belief – widespread in early modern Europe – that the ancients had kept the method of analysis hidden and had published only the rigorous synthesis, either because they considered the former not wholly demonstrative, or because they wanted to hide the method of discovery.

Another distinction which was of momentous importance for early modern mathematicians is that between problems and theorems. A problem asks a construction for its solution. It starts from certain elements considered as already constructed either by postulate or by previously established constructions. Such elements are the 'givens' (in Latin the 'data') of the problem. A problem ends with a 'Q.E.I.' or with a 'Q.E.F.' ('quod erat inveniendum' – 'what was to be discovered'–, and 'quod erat faciendum' – 'what was to be done'–, respectively). A theorem asks for a deductive proof, a sequence of propositions each following from the previous one by allowed inference rules. The starting point of the deductive chain can be either axioms or previously proved theorems. A theorem ends with 'Q.E.D.' ('quod erat demonstrandum'– 'what was to be demonstrated'). According to Pappus, therefore, there are two kinds of analysis: the former referred to problems, the latter to theorems. But it is clear from classical sources that the most important, or at least the most practiced kind, was problematic analysis: and indeed early modern European mathematicians were mainly concerned with the analysis of geometrical problems.

Another powerful idea that began to circulate in Europe at the end of the seventeenth century was that the analysis of the Greeks was not geometrical but rather symbolical: i.e. the Greeks were supposed to have had algebra and to have applied it to geometrical problem solving. The evidence that symbolic algebra was within the reach of the ancients was provided by a far from philological reading of the work of Diophantus and of parts of Euclid's *Elements*. The approach of Renaissance culture towards the classics, in sculpture, architecture, music, philosophy, and so on, was characterized by admiration united to a desire to restore the forgotten conquests of the ancients. This approach often confined with worship, a conviction of the occurrence of a decay from a glorious, golden past. The works of Euclid, Apollonius, Archimedes were considered unsurpassable models by many Renaissance mathematicians. The question that often emerged was: how could the Greeks have achieved such a wealth

of results? In the decades following the publication of the *Collection* the belief in the existence of a lost, or hidden, 'Treasure of analysis' promoted many efforts aimed at 'restoring' the ancients' method of discovery. Not everybody trod in the steps of the classicists. Typically, many promoters of the new symbolic algebra were proud to define themselves as innovators, rather than as restorers. It was common, however, even among creative algebraists such as François Viète, John Wallis and Isaac Newton, to relate symbolic algebra to the ancient analysis, to the hidden problem solving techniques of the ancients.

## 3. Descartes' method of problem solving and problem construction

Newton was deeply embedded in the conceptual space defined by Pappus and by his readers, interpreters and critics. Mainly he referred his views on mathematical method to Descartes' *Géométrie* (1637), an early source of inspiration for him and soon a target of his fierce criticisms ([5]). From this tradition Newton derived the idea that a problem, once analyzed (resolved), must be synthesized (composed or constructed).

How did Descartes define his canon of problem solving and the role of algebra in the analysis and synthesis of geometrical problems? The historian who has done most to clarify this issue is Henk Bos. It is to his work that we now turn for advice ([6]).

In book 1 of the *Géométrie* Descartes explained how one could translate a geometric problem into an equation. Descartes was able to do so by a revolutionary departure from tradition. In fact he interpreted algebraic operations as closed operations on segments. For instance, if $a$ and $b$ represent lengths of segments the product $ab$ is not conceived by Descartes as representing an area but rather another length. As he wrote: 'it must be observed that by $a^2$, $b^3$, and similar expressions, I ordinarily mean any simple lines', while before the *Géométrie* such expressions represented an area and a volume respectively (see Figure 3).

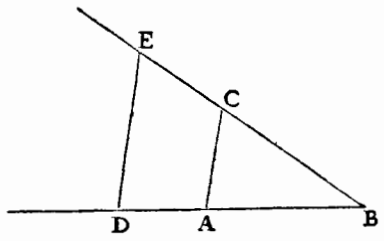
Descartes' interpretation of algebraic operations was indeed a gigantic innovation, but he proceeded wholly in line with Pappus' method of analysis and synthesis, to which he explicitly referred. In fact, according to Descartes, one has – following Pappus' prescriptions– to 'start by assuming that the problem was solved and consider a figure incorporating the solution'.[3] The segments in the figure are then denoted by letters, $a, b, c, \ldots$, for segments which are given, $x, y, z, \ldots$, for segments which are unknown. Geometrical relationships holding between the segments are then translated into corresponding equations. It is thus that one obtains a system of equations which symbolically express the assumption that the problem is solved. In fact, here we are at the very beginning of the analytic process: the unknown segments are treated as if they were known and manipulated in the equations on a par with the givens of the problem. The resolution of the equation allows the expression of the unknown $x$ in terms of given segments. We have thus moved from the assumption that the problem is solved

[3] [6] on p. 303.

298 LA GEOMETRIE.

eſt a l'autre, ce qui eſt le meſme que la Diuiſion; ou enfin trouuer vne, ou deux, ou pluſieurs moyennes proportionnelles entre l'vnité, & quelque autre ligne; ce qui eſt le meſme que tirer la racine quarrée, ou cubique, &c. Et ie ne craindray pas d'introduire ces termes d'Arithmetique en la Geometrie, affin de me rendre plus intelligibile.

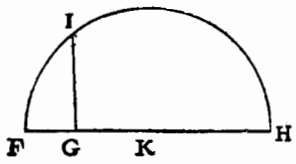
La Multiplication.




Soit par exemple A B l'vnité, & qu'il faille multiplier B D par B C, ie n'ay qu'a ioindre les poins A & C, puis tirer D E parallele a C A, & B E eſt le produit de cete Multiplication.

La Diuiſion. Ou bien s'il faut diuiſer B E par B D, ayant ioint les poins E & D, ie tire A C parallele a D E, & B C eſt le produit de cete diuiſion.

l'Extraction dela racine quarrée.




Ou s'il faut tirer la racine quarrée de G H, ie luy adiouſte en ligne droite F G, qui eſt l'vnité, & diuiſant F H en deux parties eſgales au point K, du centre K ie tire le cercle F I H, puis eſleuant du point G vne ligne droite iuſques à I, à angles droits ſur F H, c'eſt G I la racine cherchée. Ie ne dis rien icy de la racine cubique, ny des autres, à cauſe que i'en parleray plus commodement cy aprés.

Commẽt on peut Mais ſouuent on n'a pas beſoin de tracer ainſi ces ligne

4

Figure 3. Descartes' geometric interpretation of algebraic operations. He writes: 'For example, let $AB$ be taken as unity, and let it be required to multiply $BD$ by $BC$. I have only to join the points $A$ and $C$, and draw $DE$ parallel to $CA$; and then $BE$ is the product of $BD$ and $BC$'. So, given a unit segment, the product of two segments is represented by another segment, not by a surface. The second diagram is the construction of the square root of $GH$. Given $GH$ and a unit segment $FG$, one draws the circle of diameter $FG + GH$ and erects $GI$, the required root. Source: [5], 4.

(the first step of the analysis) to a reduction of the unknown, sought magnitude to the givens. This is why Descartes, and the other early-modern promoters of algebra, associated algebra with the method of analysis.

The *resolution* of the equation is not, however, the *solution* of the problem. In fact, the solution of the problem must be a geometrical construction of the sought magnitude in terms of legitimate geometrical operations performed on the givens ('Q.E.F.'!). We now have to move from algebra back to geometry again. Descartes understood this process from algebra to geometry as follows: the real roots of the equation (for him if there are no real roots, then the problem admits no solution) must be geometrically constructed. After Descartes, this process was known as the 'construction of the equation'. This is where the synthetic, compositive part of the whole process begins.

Descartes accepted from tradition the idea that such constructions must be performed by intersection of curves. That is to say, the real roots are geometrically represented by segments, and such segments are to be constructed by intersection of curves. As a matter of fact, the construction of the equation presented the geometer with a *new* problem: not always an easy one. One had to choose two curves, position and scale them, such that their intersections determine points from which segments – whose lengths geometrically represent the roots of the equation – can be drawn (see Figure 4).

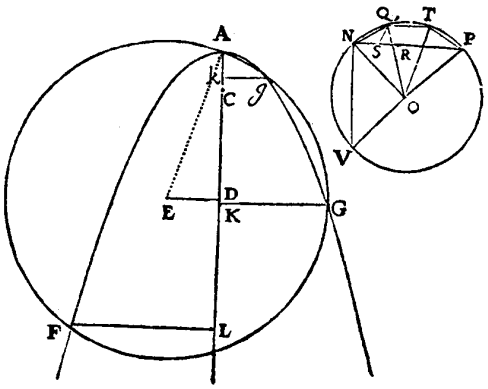



Figure 4. Construction of a third-degree equation in Descartes' *Géométrie*. The problem of trisecting angle $NOP$ is *resolved* ('resolutio' is the Latin translation of the Greek 'analysis') by a third-degree equation. Descartes *constructs* the roots ('constructio' or 'compositio' translate 'synthesis') via intersection of circle and parabola. The segments $kg$, $KG$ and $LF$ represent two positive and one negative root. The smaller of the two positive roots $kg$ must be 'taken as the length of the required line $NQ$'. $KG$ is equal to $NV$, 'the chord subtended by one-third the arc $NVP$'. Source: [5], 208.

The synthetic part of Descartes' process of problem-solving gave rise to two questions: which curves are admissible in the construction of equations? which curves,

among the admissible, are to be preferred in terms of simplicity? In asking himself these questions Descartes was continuing a long debate concerning the role and classification of curves in the solution of problems. A tradition that, once again, stems from Pappus, and the interpretations of Pappus given by mathematicians such as Viète, Ghetaldi, Kepler, and Fermat. His answer was that only 'geometrical curves' (we would say 'algebraic curves') are admissible in the construction of the roots of equations and that one has to choose the curves of the lowest possible degree as these are the simplest. Descartes instead excluded 'mechanical curves' (we would say transcendental curves) as legitimate tools of construction.

Notice that Descartes presented his canon of problem resolution and construction in aggressively anti-classicist terms. His algebraic method, he claimed, was superior to the ones followed by the ancients. He gave pride of place to a problem discussed in Pappus' *Mathematical Collection* that – according to Descartes – neither Euclid nor Apollonius could solve. He proudly showed to the readers of the slim *Géométrie* that, by applying algebra to geometry, he could easily achieve a solution not included in the ponderous Pappusian tomes.[4]

## 4. Newton versus Descartes

Newton sharply criticized Descartes' canon of problematic analysis and construction.[5] Newton's point was that geometrical constructions have to be carried on in terms independent from algebra. Newton elaborated his criticism to Descartes in his Lucasian Lectures on Algebra which were held before 1684 and which, in somewhat modified form, appeared in 1707 as *Arithmetica Universalis* ([8], vol. 2, 3–135). The *Arithmetica Universalis* ends with an Appendix devoted to the 'construction of equations' which abounds with oft-quoted statements in favour of pure geometry and against the 'Moderns' (read Descartes) who have lost the 'Elegance' of geometry:

> Geometry was invented that we might expeditiously avoid, by drawing Lines, the Tediousness of Computation. Therefore these two sciences [Geometry and Arithmetical Computation] ought not be confounded. The Ancients did so industriously distinguish them from one another, that they never introduced Arithmetical Terms into Geometry. And the Moderns, by confounding both, have lost the Simplicity in which all the Elegance of Geometry consists.[6]

[4] Briefly said, Pappus problem requires the determination of the locus of points $P$ such that their distances $d_i$ $(i = 1, 2, 3, 4)$ from four lines given in position are such that $d_1 d_2 = k(d_3 d_4)$. In the *Géométrie* Descartes introduces a system of oblique coordinates, and notices that the distance of a point from a line is given by an expression of the form $ax + by + c$. Therefore Pappus 4-lines locus has a second-degree defining equation: namely it will be a conic section. The algebraic approach immediately allowed Descartes to generalize Pappus problem for any number of lines.

[5] Further information on Newton's criticisms to Descartes can be gained from [7].

[6] [8], vol. 2, 228.

Such statements have often puzzled commentators since they occur in a work devoted to algebra and in which the advantage of algebraic analysis is displayed in a long section on the resolution of geometrical problems. Why was Newton turning his back to 'arithmetic'[7] now saying that algebra and geometry should be kept apart? In order to understand this seemingly paradoxical position we have to briefly recall that according to Descartes the demarcation between admissible and inadmissible curves as means of construction was that between geometrical and mechanical curves. Ultimately, Descartes was forced to make recourse to algebraic criteria of demarcation and simplicity: in fact, algebraic curves coincided for him with the loci of polynomial equations, and the degree of the equation allowed him to rank curves in terms of their simplicity.

As far as demarcation is concerned, in the *Arithmetica Universalis* Newton maintained that it would be wrong to think that a curve can be accepted or rejected in terms of its defining equation. He wrote:

> It is not the Equation, but the Description that makes the Curve to be a Geometrical one. The Circle is a Geometrical Line, not because it may be expressed by an Equation, but because its Description is a Postulate.[8]

Further, Descartes' classification of curves in function of the degree of the equation – Newton claimed – is not relevant for the geometrician, who will choose curves in function of the simplicity of their description. Newton, for instance, observed that the equation of a parabola is simpler than the equation of the circle. However, it is the circle which is simpler and to be preferred in the construction of problems:

> It is not the simplicity of its equation, but the ease of its description, which primarily indicates that a line is to be admitted into the construction of problems. […] On the simplicity, indeed, of a construction the algebraic representation has no bearing. Here the descriptions of curves alone come into the reckoning.[9]

Newton observed that from this point of view, the conchoid, a fourth degree curve, is quite simple. Independently of considerations about its equation, its mechanical description – he claimed – is one of the simplest and most elegant; only the circle is simpler. Descartes' algebraic criterion of simplicity is thus deemed alien to the constructive, synthetical, stage of problem solving. The weakness of Newton's position is that the concepts of simplicity of tracing, or of elegance, to which he continuously refers are qualitative and subjective. One should be aware that no compelling reason is given in support of Newton's evaluations on the simplicity of his preferred constructions: his are largely aesthetic criteria. Considering them is however crucial for our understanding of Newton's views concerning mathematical method.

[7] Notice that Newton employed the term 'universal arithmetic' for algebra, since it is concerned with the doctrine of operations, not applied to numbers, but to general symbols.

[8] [8], vol. 2, 226.

[9] [9], vol. 5, 425–7.

As a matter of fact, Newton – this master of algebraic manipulations – in the mid 1670s developed a deep distaste for symbolism and distanced himself from the mathematics of the 'moderns'. He wrote:

> The Modern Geometers are too fond of the Speculation of Equations. The Simplicity of these is of an Analytick Consideration. [in the Appendix to the *Arithmetica Universalis*] [w]e treat of Composition, and Laws are not given to Composition from Analysis. Analysis does lead to Composition: but it is not true Composition before it is freed from Analysis. If there be never so little Analysis in Composition, that Composition is not yet real. Composition in it self is perfect, and far from a Mixture of Analytick Speculations.[10]

This position, let me restate it, does not exclude the use of algebra in the analysis; it does, however, rule out algebraic criteria of demarcation and simplicity from the synthesis. As Newton was to affirm in a manuscript dating from the early 1690s:

> if a question be answered […] that question is resolved by the discovery of the equation and composed by its construction, but it is not solved before the construction's enunciation and its complete demonstration is, with the equation now neglected, composed.[11]

But, around 1680, Newton moved a step forward in his opposition to the method proposed in the *Géométrie*: not only Cartesian synthesis, but also Cartesian analysis fell under his fierce attack. He developed a deep admiration for the ancient Greek mathematicians, while he criticized in bitter terms the symbolical analysis pursued by the moderns. He began to doubt that the analysis of the Greeks was algebraical, he rather suspected that Euclid and Apollonius possessed a more powerful *geometrical analysis* displayed in the three lost books on *Porisms* attributed to Euclid and described in Book 7 of the *Mathematical Collection*. So not only the composition (the synthesis) had to be freed from algebra, the algebraic calculus had to be avoided also in the process of resolution (the analysis). His target was often Descartes. For instance in the late 1670s, commenting on Descartes' solution of Pappus problem, he stated with vehemence:

> To be sure, their [the Ancients'] method is more elegant by far than the Cartesian one. For he [Descartes] achieved the result by an algebraic calculus which, when transposed into words (following the practice of the Ancients in their writings), would prove to be so tedious and entangled as to provoke nausea, nor might it be understood. But they accomplished it by certain simple propositions, judging that nothing written in a different style was worthy to be read, and in consequence concealing the analysis by which they found their constructions.[12]

[10] [8], vol. 2, 250.
[11] [9], vol. 7, 307.
[12] [9], vol. 4, 277.

Newton was not alone in his battle against the algebraists. Similar statements can be found in the polemic works of Thomas Hobbes. But probably the deepest influence on Newton in this matter was played by his mentor Isaac Barrow. Newton's quest for the ancient, non-algebraical, porismatic analysis led him to develop an interest in projective geometry (see Figure 5).

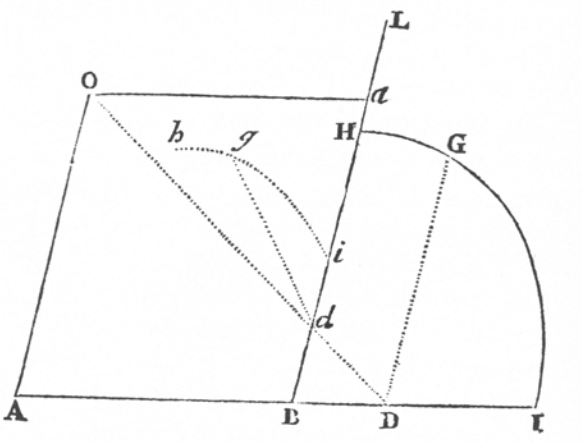



Figure 5. Newton was interested in using projective transformations as a heuristic analytic tool. Here we reproduce the diagram for Lemma 22, Book 1, of the *Principia*. In this Lemma we are taught how 'To change figures into other figures of the same class' (namely, algebraic curves of the same degree). The figure to be transmuted is the curve $HGI$. Draw the straight parallel lines $AO$ and $BL$ cutting any given third line $AB$ in $A$ and $B$. Then from some point $O$ in the line $AO$ draw the straight line $OD$. From the point $d$ erect the ordinate $dg$ (you can choose any angle between the 'new ordinate' $dg$ and the 'new abscissa' $ad$). The new ordinate and abscissa have to satisfy the following conditions: $AD = (AO \times AB)/ad$ and $DG = (AO \times dg)/ad$. These transformations are exactly those occurring between figures projected from one plane into another. Now suppose that point $G$ 'be running through all the points in the first figure [$HGI$] with a continual motion; then point $g$ – also with a continual motion – will run through all the points in the new figure [$hgi$]'. Source: [11], 162.]

He convinced himself that the ancients had used projective properties of conic sections in order to achieve their results. Moving along these lines he classified cubics into five projective classes.[13]

[13] From his work on cubics ([8], vol. 2, 137–161) Newton derived two lessons. First, Descartes' classification of curves by degree is an algebraic criterion which has little to do with simplicity. Indeed, cubics have rather complex shapes compared to mechanical (transcendental) curves such as the Archimedean spiral. Second, it is by making recourse to projective classification that one achieves order and generality.

## 5. Newton's new analysis

Now that we know more about Newton's views concerning the role of algebraic symbolism in the method of problem solving, we are in the position to step back to Newton's memorandum on his early mathematical discoveries that I quoted in Section 1. There he mentions the direct and the inverse methods of fluxions. The direct method allowed the determination of tangents (and curvature) to plane curves. Newton approached this problem by conceiving curves as generated by the continuous 'flow' of a point. He called the geometric magnitudes generated by motion 'fluents', while 'fluxions' are the instantaneous rates of flow. In the 1690s he denoted fluxions with overdots, so that the fluxion of $x$ is $\dot{x}$. He deployed a variety of strategies in order to determine tangents. Some of them are algorithmic, but in many cases Newton made recourse to kinematic methods. In Newton's mathematical writings the algorithm is indeed deeply intertwined with geometrical speculations.

By resolving motion into rectilinear components Newton could determine the tangent by composition of motions, even in the case of mechanical lines (see Figure 6). Indeed, the possibility to deal with transcendental curves (as the spiral and the cycloid) was top in Newton's agenda. Or one could focus attention on the 'moment of the arc' generated in a very short interval of time (Newton termed the infinitesimal increment acquired in an infinitesimal interval of time a 'moment') and establish a proportion between the moment of the abscissa and the moment of the ordinate and other finite lines embedded in the figure. When the curve was expressed symbolically via an equation Newton had 'rules' which allowed him to calculate the tangent (see Figure 7). One recognizes here rules which are 'equivalent' to those of the differential calculus; but the reader should be reminded that this equivalence was, and still is, object of debate.

The inverse method of fluxions was Newton's masterpiece. It is this method that allowed him to approach the problem of 'squaring curves'. By conceiving a surface $t$ as generated by the flow of the ordinate $y$ which slides at a right angle over the abscissa $z$, he understood that the rate of flow of the surface's area is equal to the ordinate (he stated $\dot{t}/\dot{z} = y/1$). This is how the idea of integration as anti-differentiation was born in Newton's mind. His approach consisted in applying the direct method to 'equations at will [which] define the relationship of $t$ to $z$'. One thus obtains an equation for $\dot{t}$ and $\dot{z}$, and so 'two equations will be had, the latter of which will define the curve, the former its area'.[14] Following this strategy Newton constructed a 'Catalogue of curves' which can be squared by means of 'finite equations' (see Figure 8). In Leibnizian terms, he built the first integral tables in the history of mathematics.

Newton attached much importance to the inverse method. With almost visionary mathematical understanding of what is truly revolutionary, while still in his early years, he wrote:

> If two Bodys $A$ & $B$, by their velocitys $p$ & $q$ describe y$^{e}$ lines $x$ & $y$.

[14] [9], vol. 3, 197.

& an Equation bee given expressing y$^{e}$ relation twixt one of y$^{e}$ lines $x$, & ye ratio $q/p$ of their motions $q$ & $p$; To find y$^{e}$ other line $y$. Could this ever bee done all problems whatever might bee resolved.[15]

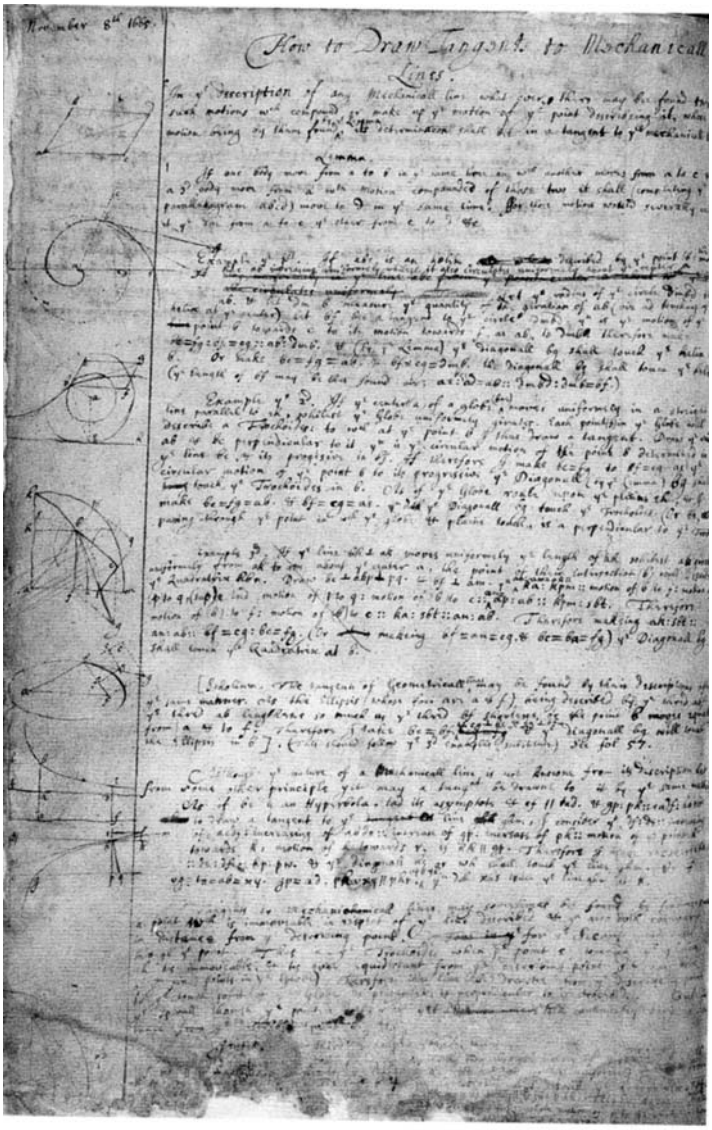
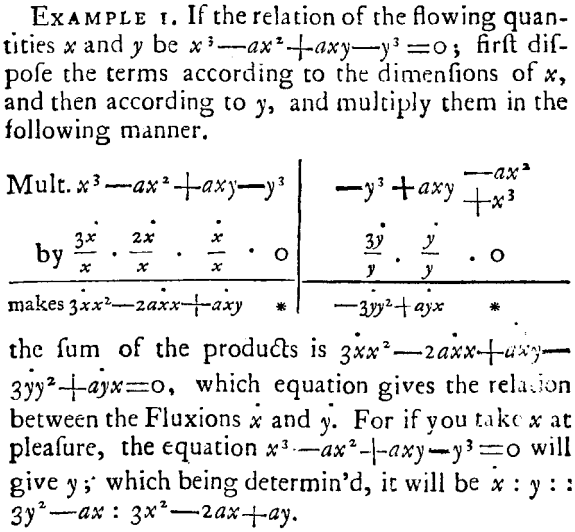



Figure 6. Newton's early work (November 1666) on tangents to 'mechanicall lines' (i.e. transcendental plane curves). His technique consisted in conceiving curves as generated by motion and resolving motion into components. Source: [9], vol. 1, 378.

[15] [9], vol. 1, 403.



EXAMPLE 1. If the relation of the flowing quantities $x$ and $y$ be $x^3 - ax^2 + axy - y^3 = 0$; firſt diſpoſe the terms according to the dimenſions of $x$, and then according to $y$, and multiply them in the following manner.

Mult. $x^3 - ax^2 + axy - y^3$ | $-y^3 + axy \frac{-ax^2}{+x^3}$

by $\frac{3\dot{x}}{x} \cdot \frac{2\dot{x}}{x} \cdot \frac{\dot{x}}{x} \cdot 0$ | $\frac{3\dot{y}}{y} \cdot \frac{\dot{y}}{y} \cdot 0$

makes $3\dot{x}x^2 - 2a\dot{x}x + a\dot{x}y$ * | $-3\dot{y}y^2 + a\dot{y}x$ *

the ſum of the products is $3\dot{x}x^2 - 2a\dot{x}x + a\dot{x}y - 3\dot{y}y^2 + a\dot{y}x = 0$, which equation gives the relation between the Fluxions $\dot{x}$ and $\dot{y}$. For if you take $x$ at pleaſure, the equation $x^3 - ax^2 + axy - y^3 = 0$ will give $y$; which being determin'd, it will be $\dot{x} : \dot{y} :: 3y^2 - ax : 3x^2 - 2ax + ay$.

Figure 7. Newton's algorithm for the direct method of fluxions. In this example he calculates the relation between fluxions (instantaneous speeds) $\dot{x}$ and $\dot{y}$ of fluent quantities (magnitudes changing continuously in time) $x$ and $y$ related by the equation $x^3 - ax^2 + axy - y^3 = 0$. Source: [8], vol. 1, 50.

In this context Newton developed techniques equivalent to integration by parts and substitution.

Newton labelled the techniques of series expansion, tangent determination and squaring of curves as the 'method of series and fluxions'. This was, he proudly stated, a 'new analysis' which extended itself to objects that Descartes had banished from his 'common analysis' – such as mechanical curves – thanks to the use of infinite series:

> And whatever common analysis performs by equations made up of a finite number of terms (whenever it may be possible), this method may always perform by infinite equations: in consequence, I have never hesitated to bestow on it also the name of analysis.[16]

According to Newton, the 'limits of analysis are enlarged by [...] infinite equations: [...] by their help analysis reaches to all problems'.[17]

[16] [9], vol. 2, 241.
[17] [10]

*A* TABLE *of ſome Curves related to Rectilinear Figures, conſtructed by* PROBLEM VII.

| | Order of Curves. | Values of the Areas. |
|---|---|---|
| I | $dz^{n-1}=y$ | $\frac{d}{n}z^n=t$ |
| II | $\frac{dz^{n-1}}{ee+2efz^n+ffz^{2n}}=y$ | $\frac{dz^n}{ne^2+nefz^n}=t$ |
| III 1 | $dz^{n-1}\sqrt{e+fz^n}=y$ | $\frac{2d}{3nf}R^3=t$ |
| III 2 | $dz^{2n-1}\sqrt{e+fz^n}=y$ | $\frac{-4e+6fz^n}{15nff}dR^3=t$ |
| III 3 | $dz^{3n-1}\sqrt{e+fz^n}=y$ | $\frac{16ee-24efz^n+30ffz^{2n}}{105nf^3}dR^3=t$ |
| III 4 | $dz^{4n-1}\sqrt{e+fz^n}=y$ | $\frac{-96e^3+144e^2fz^n-180ef^2z^{2n}+210f^3z^{3n}}{945nf^4}dR^3=t$ |
| IV 1 | $\frac{dz^{n-1}}{\sqrt{e+fz^n}}=y$ | $\frac{2d}{nf}R=t$ |
| IV 2 | $\frac{dz^{2n-1}}{\sqrt{e+fz^n}}=y$ | $\frac{-4e+2fz^n}{3nff}dR=t$ |
| IV 3 | $\frac{dz^{3n-1}}{\sqrt{e+fz^n}}=y$ | $\frac{16e^2-8efz^n+6ffz^{2n}}{15nf^3}dR=t$ |
| IV 4 | $\frac{dz^{4n-1}}{\sqrt{e+fz^n}}=y$ | $\frac{-96e^3+48e^2fz^n-36ef^2z^{2n}+30f^3z^{3n}}{105nf^4}dR=t$ |

*pag.* 137.

Figure 8. The beginning of Newton's table of curves (an integral table, in Leibnizian terms), obtained thanks to understanding of what we call the 'fundamental theorem of calculus'. Here Newton lists the first four 'orders'. $z$ is the abscissa, $y$ the ordinate, $t$ the area. In Newton's notation $\dot{t}/\dot{z} = y/1$. Notice that $d, e, f, g, h$ are constants ($d$ is a constant!), $\eta$ is integer or fractional, and $R$ stands for $\sqrt{e+fz^\eta}$ or $\sqrt{e+fz^\eta+gz^{2\eta}}$. Source: [8], vol. 1, 105.

## 6. Newton's synthetical method

One should recall that the 'new analysis' occupied in Newton's agenda a place which, according to the Pappusian canon, was subsidiary to the synthesis or construction, and that the construction had to be carried on in terms independent of algebraic criteria. For instance, as to the squaring of curves (in Leibnizian terms, integration) he wrote:

> After the area of some curve has thus been found, careful considerations should be given to fabricating a demonstration of the construction which

> as far as permissible has no algebraic calculation, so that the theorem embellished with it may turn out worthy of public utterance.[18]

Newton therefore devoted great efforts to providing geometrical demonstrations, somewhat reminiscent of Archimedean exhaustion techniques, of his 'analytical' quadratures. Only such demonstrations were deemed by him 'worthy of public utterance'.

It is in this context that Newton in the 1670s began reworking his early discoveries in 'new analysis' in terms that he conceived concordant with the constructive geometrical methods of the ancients. He termed this more rigorous approach the 'synthetical method of fluxions' and codified it around 1680 in a treatise entitled *Geometria curvilinea* ([9], vol. 4, 420–521). In this method no infinitesimals, or 'moments', occurred and no algebraic symbols were deployed. Everything was based upon geometric limit procedures that Newton termed the 'method of first ratios of nascent quantities and last ratios of vanishing quantities'. It is this method that was widely deployed in the *Principia* (1687) (see Figure 9). It is somewhat astonishing to see one of the most

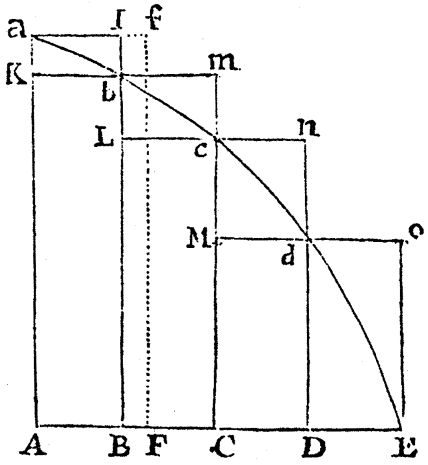



Figure 9. In Section 1, Book 1 of the *Principia* Newton lays down his 'method of first and last ratios', a geometric limit procedure that allows him to avoid infinitesimals. In Lemma 2 Newton shows that a curvilinear area $AabcdE$ can be approached as the limit of inscribed $AKbLcMdD$ or circumscribed $AalbmcndoE$ rectilinear areas. Each rectilinear surface is composed of a finite number of rectangles with equal bases $AB$, $BC$, $CD$, etc. The proof is magisterial in its simplicity. Its structure is still retained in present day calculus textbooks in the definition of the Riemann integral. It consists in showing that the difference between the areas of the circumscribed and the inscribed figures tends to zero, as the number of rectangles is 'increased *in infinitum*'. In fact this difference is equal to the area of rectangle $ABla$ which, 'because its width $AB$ is diminished *in infinitum*, becomes less than any given rectangle'. In Newton's terms $AB$ is a 'vanishing quantity'. Source: [11], 74.

[18] [9], vol. 3, 279.

creative algebraists of the history of mathematics spend so much time and effort in reformulating his analytical results in geometric terms, but Newton had compelling reasons to do so.

First, Newton in his programme of reformation of natural philosophy attributed an important role to mathematics as a source of certainty. From the early 1670s he expressed his distaste for the probabilism and hypotheticism that was characteristic of the natural philosophy[19] practiced at the Royal Society by people like Robert Hooke and Robert Boyle. His recipe was to inject mathematics into natural philosophy. As he stated:

> by the help of philosophical geometers and geometrical philosophers, instead of the conjectures and probabilities that are being blazoned about everywhere, we shall finally achieve a science of nature supported by the highest evidence. ([12])

But if mathematics has to provide certainty to natural philosophy her methods must be above dispute, and Newton was keenly aware of the fact that the new analysis was far from being rigorous.

Second, Newton soon developed a deep anti-Cartesianism associated with a conviction of the superiority of the ancients over the moderns. From his point of view Descartes was the champion of an impious mechanistic philosophy which, conceiving nature as an autonomous mechanism, denied any role to God's providence. Newton conceived himself as a restorer of an ancient, forgotten philosophy according to which nature is always open to the providential intervention of God. Indeed, he thought that, according to the theory of gravitation – which he was convinced the ancient Hebrews possessed–, the quantity of motion in the universe was bound to decline if divine intervention had not prevented the 'corruption of the heavens'. The modern philosophers were dangerous from a theological point of view and had to be opposed on all grounds. Therefore, also in mathematics Newton looked with admiration to ancient exemplars and conceived himself as a restorer of their glory. It goes without saying that the above reasons led Newton into a condition of strain, since his philosophical values were at odds with his mathematical practice, which was innovative, symbolical, and – pace Newton – deeply Cartesian.

Several hitherto unexplained aspects of Newton's mathematical work are related to this condition of stress and strain that characterizes his thoughts on mathematical method. Why did Newton fail to print his method of series and fluxions before the inception of the priority dispute with Leibniz? Why did he hide his competence in quadratures when writing the *Principia*, which are written mostly in geometrical style? Even though there is no single answer to these vexed questions, I believe that Newton's conviction that the analytical symbolical method is only a heuristic tool,

[19] For Newton the aim of 'natural philosophy' is to deduce the forces from phenomena established by experiment, and – once established the forces – to deduce new phenomena from them. Nowadays we would call this enterprise 'physics'.

not 'worthy of public utterance', can in part explain a policy of publication which was to have momentous consequences in the polemic with Leibniz.

## 7. Leibniz's views

When the war with Leibniz exploded in 1710 Newton had to confront an opponent who not only advanced mathematical results equivalent to his, but was promoting a different view concerning mathematics.[20]

The rhetoric on the novelty of the calculus pervades Leibniz's writings. Reference to the ancient mathematicians generally took the rather abused form of a tribute to Archimedes' 'method of exhaustion'. Leibniz in most of his declarations concerning the calculus wished to highlight the novelty and the revolutionary character of his algorithm, rather than continuity with ancient exemplars. This approach is quite at odds with Newton's 'classicism'. Furthermore, Leibniz often referred to the heuristic character of the calculus understood as an algorithm independent from geometrical interpretation. It is exactly this independence that would render the calculus so efficacious in the process of discovery. The calculus, according to Leibniz, should also be seen as an *ars inveniendi* (an art of discovery): as such it should be valued by its fruitfulness, rather than by its referential content. We can calculate, according to Leibniz, with symbols devoid of referential content (for instance, with $\sqrt{-1}$), provided the calculus is structured in such a way as to lead to correct results.[21]

Writing to Christiaan Huygens in September 1691, Leibniz affirmed with pride:

> It is true, Sir, as you correctly believe, that what is better and more useful in my new calculus is that it yields truths by means of a kind of analysis, and without any effort of the imagination, which often works as by chance. ([13])

[20] The circumstances surrounding the controversy between Newton and Leibniz have been analysed in detail by Rupert Hall [15] and Tom Whiteside [9], vol.8. In broad outlines let me recall a few bare facts. Newton formulated his method of series and fluxions between 1665 and 1669. Leibniz had worked out the differential and integral calculus around 1675 and printed it in a series of papers from 1684. It is clear from manuscript evidence that he arrived at his results independently from Newton. It is only in part in Wallis' *Algebra* in 1685 and *Works* in 1693 and 1699, and in full in an appendix to the *Opticks* in 1704, however, that Newton printed his method. In 1710 a British mathematician, John Keill, stated in the *Philosophical Transactions of the Royal Society* that Leibniz had plagiarized Newton. After Leibniz's protest a committee of the Royal Society secretly guided by its President, Isaac Newton, produced a publication – the so-called *Commercium epistolicum* (1713) – in which it was maintained that Newton was the 'first inventor' and that '[Leibniz's] Differential Method is one and the same with [Newton's] Method of Fluxions'. It was also suggested that Leibniz, after his visits to London in 1673 and 1676, and after receiving letters from Newton's friends, and from Newton himself (in fact Newton addressed two letters to Leibniz in 1676) had gained sufficient information about Newton's method to allow him to publish the calculus as his own discovery, after changing the symbols. It is only after the work of historians such as Fleckenstein, Hofmann, Hall and Whiteside that we have the proof that this accusation was unjust. Newton and Leibniz arrived at equivalent results independently and following different paths of discovery.

[21] Complex numbers received a geometric interpretation only around 1800 thanks to Jean Robert Argand, Carl Friedrich Gauss, and Caspar Wessel.

Leibniz was thus praising the calculus as a *cogitatio caeca* and promoted the 'blind use of reasoning' among his disciples. Nobody, according to Leibniz, could follow a long reasoning without freeing the mind from the 'effort of imagination'.[22]

Leibniz conceived of himself as the promoter of new methods of reasoning, rather than 'just' a mathematician. The calculus was just one successful example of the power of algorithmic thinking. The German diplomat was interested in promoting in Europe the formation of a group of intellectuals who could extend a universal knowledge achieved thanks to a new algorithm that he termed *universal characteristic*. He thus helped to form a school of mathematicians who distinguished themselves by their ability in handling the differentials and the integrals and by their innovative publication strategy. Thanks to Leibniz's recommendation, they colonized chairs of mathematics all over Europe. The efficacy of this new algorithm was affirmed to be independent from metaphysical or cosmological questions. The persons who practised it had to be professional mathematicians, rather than 'geometrical philosophers', able to teach and propagate knowledge of calculus.

A typical Leibnizian attitude emerges in the context of the vexed question of the existence of infinitesimals. The new calculus was often attacked, since – it was maintained – it employed symbols devoid of meaning, such as differentials ordered into a bewildering hierarchy of orders. Newton, as we know, was particularly sensitive to such criticisms, and tried in his synthetical method to dispense with infinitely small quantities. Leibniz, on the other hand, repeated many times that for him the question of the existence of infinitesimals had to be distinguished from that of their usefulness as algorithmic devices. While he was leaning, for philosophical reasons, towards a denial of the existence of infinitesimals, he also wanted to stress that this ontological question was somewhat extraneous to mathematics. A typical statement, written in the early years of the eighteenth century, is the following:

> We have to make an effort in order to keep pure mathematics chaste from metaphysical controversies. This we will achieve if, without worrying whether the infinites and infinitely smalls in quantities, numbers and lines are real, we use infinites and infinitely smalls as an appropriate expression for abbreviating reasonings. ([14])

Leibniz was thus leaving to his disciples the choice of maintaining, *philosophically speaking*, different approaches to the ontological question on the existence of infinitesimals. What he wished to defend was their utility as symbols in mathematical calculation.

## 8. The war against Leibniz: methodological aspects

When Newton had to confront Leibniz in the squabble over priority he was concerned in building up a forensic and historical document whose purpose was to prove

[22] [14], 205.

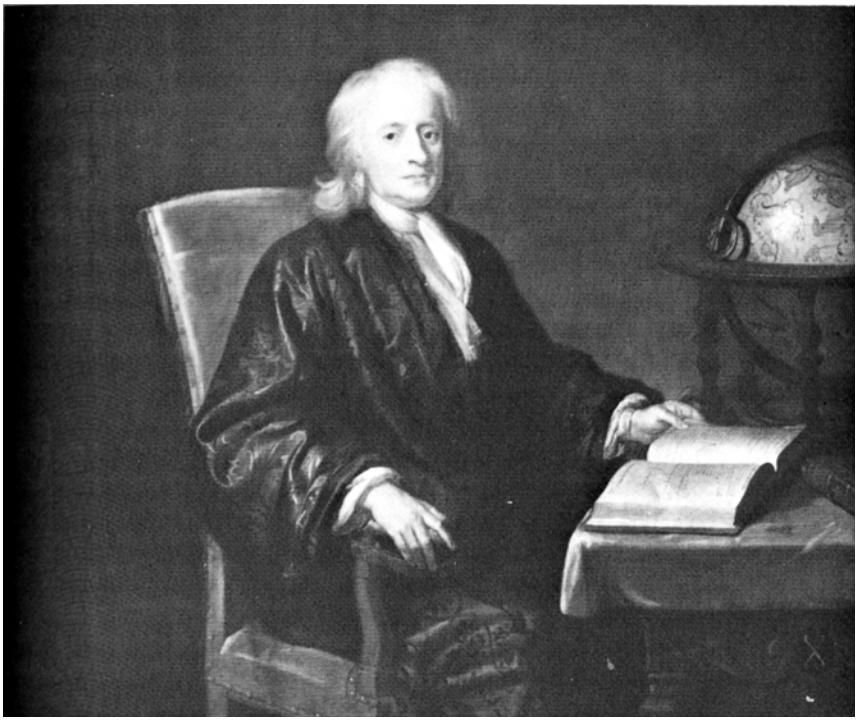


Figure 10. A portrait of Newton in old age (Source: [1], 831). He proudly opens the *Principia* at a page devoted to the attraction of extended bodies. In dealing with this problem Newton made recourse to his 'inverse method of fluxions' (the equivalent of Leibniz's integral calculus) which allowed him to 'square curves'. As a matter of fact, only by making recourse to his tables of curves ('integral tables'), see Figure 8, could Newton solve several problems in the *Principia*. Such analytic methods were not, however, made explicit to the reader. In the polemic with the Leibnizians – who claimed that absence of calculus from the *Principia* was proof positive of Newton's ignorance of quadrature techniques prior to 1687 – Newton was forced to maintain, with some exaggeration, that 'By the help of this new Analysis Mr Newton found out most of the Propositions in his *Principia Philosophiae*. But because the Ancients for making things certain admitted nothing into Geometry before it was demonstrated synthetically, he demonstrated the Propositions synthetically that the systeme of the heavens might be founded upon good Geometry. And this makes it now difficult for unskillful men to see the Analysis by w$^{ch}$ those Propositions were found out.' ([9], vol. 8, 599). On the issue of Newton's use of analytic methods in the *Principia* see [16].

Leibniz's plagiarism. But he did not do only this, he also wished to highlight the superiority of his *method* over Leibniz's *calculus*. The mathematical programme that Leibniz was promoting with so much success was at odds with Newton's deeply felt values.

There is not only mathematics in this story, of course. Leibniz had to be opposed for a series of reasons that have to do with the Hannoverian succession. The German,

in fact, who was employed by the Hannover family, wished to move to London as Royal Historian. The idea of having in England such a towering intellectual who was defending a philosophical view which contradicted Newton's voluntarist theology and who was promoting the unification of the Christian Churches was anathema for Newton and his supporters.

For our purposes, it is interesting to turn to some passages that Newton penned in 1715 contained in an anonymous 'Account' to a collection of letters, the *Commercium epistolicum*, that the Royal Society produced in order to demonstrate Leibniz's plagiarism.

In the 'Account', speaking of himself in the third person, Newton made it clear that Leibniz had only approached the analytical, heuristic part of the problem-solving method. He wrote:

> Mr. Newton's Method is also of greater Use and Certainty, being adapted either to the ready finding out of a Proposition by such Approximations as will create no Error in the Conclusion, or to the demonstrating it exactly; Mr. Leibniz's is only for finding it out.[23]

So according to Newton, Leibniz had achieved only the first stage of the Pappusian method and had not attained the rigorous, constructive demonstration. This, as we know, had to be carried on in purely geometric terms.

Further, Newton insisted on the fact that the emphasis with which Leibniz praised the power of his symbolism was excessive. Algorithm is certainly important for Newton, but it has to be viewed only as a component of the method:

> Mr Newton — he wrote – doth not place his Method in Forms of Symbols, nor confine himself to any particular Sort of Symbols.[24]

Finally, Newton noticed that in his method of first and last ratios no infinitesimals occur, everything being performed according to limiting procedures. From Newton's point of view the avoidance of infinitesimals and the possibility of interpreting algebraic symbols as geometric magnitudes had the double advantage of rendering his method endowed with referential content and consonant with ancient mathematics:

> We have no ideas of infinitely little quantities & therefore Mr Newton introduced fluxions into his method that it might proceed by finite quantities as much as possible. It is more natural & geometrical because founded on *primae quantitatum nascentium rationes* [first ratios of nascent quantities] w$^{ch}$ have a being in Geometry, whilst *indivisibles* upon which the Differential method is founded have no being either in Geometry or in nature. […] Nature generates quantities by continual flux or increase, & the ancient Geometers admitted such a generation of areas & solids […]. But the summing up of indivisibles to compose an area or solid was never yet admitted into Geometry.[25]

[23] Cited in [15], 296.
[24] Cited in [15], 294.

Nature and geometry are the two key concepts: they allow Newton to defend his method because of its continuity with ancient tradition as well as its ontological content.

In his polemic writings against Leibniz Newton engineered an attack which was aimed at proving the German's plagiarism. One of Newton's priorities was to assemble evidence which proved Leibniz guilty, and he did so with means that show his ability to employ archival sources as well as his prejudice and egotism. However, Newton also defended positions concerning mathematical method that have deep roots in his protracted opposition against Descartes and the 'modern mathematicians' who, by confounding geometry and algebra, 'have lost the Simplicity in which all the Elegance of Geometry consists'.

## References


[1] Add MS 3968.41 f. 85. Cited in Westfall, R. S., *Never at Rest: a Biography of Isaac Newton*. Cambridge University Press, Cambridge 1980, 143. This is the best biography of Newton.

[2] On Newton's philosophical programme one can read: Cohen, I. B., *The Newtonian Revolution*. Cambridge University Press, Cambridge 1980. Smith, G. E., The methodology of the *Principia*. In *The Cambridge Companion to Newton* (ed. by I. Bernard Cohen and George E. Smith). Cambridge University Press, Cambridge 2002, 137–173. Stein, H., Newton's metaphysics. In *The Cambridge Companion to Newton* (ed. by I. Bernard Cohen and George E. Smith). Cambridge University Press, Cambridge 2002, 256–307.

[3] Panza M., *Newton et les Origines de l'Analyse: 1664–1666*. Blanchard, Paris 2005.

[4] Pappus of Alexandria, *Book 7 of the Collection* (ed. by Alexander Jones). Sources in the History of Mathematics and the Physical Sciences 8, Springer, New York, Berlin, Heidelberg, Tokyo 1986, 82–4.

[5] Descartes, R., *The Geometry of René Descartes with a Facsimile of the First Edition* (ed. by D. E. Smith and M. L. Latham). Dover, New York 1954.

[6] Bos, H. J. M., *Redefining Geometrical Exactness: Descartes' Transformation of the Early Modern Concept of Construction*. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo 2001.

[7] Galuzzi, M., I *marginalia* di Newton alla seconda edizione latina della *Geometria* di Descartes e i problemi ad essi collegati. In *Descartes: il Metodo e i Saggi* (ed. by G. Belgioioso, G. Cimino, P. Costabel, G. Papuli), Istituto dell' Enciclopedia Italiana, Roma 1990, 387–417.

[8] Newton, I., *The Mathematical Works of Isaac Newton* (ed. by Derek T. Whiteside). 2 vols., Johnson Reprint Corp., New York, London 1964–1967.

[9] Newton, I., *The Mathematical Papers of Isaac Newton* (ed. by Derek T. Whiteside). 8 vols., Cambridge University Press, Cambridge, London, New York 1967–1981. In quoting from this work we follow Whiteside's translation from the Latin.

[10] Newton, I., *The Correspondence of Isaac Newton* (ed. by H. W. Turnbull *et al.*). 7 vols., Cambridge University Press, Cambridge, 1959–1977, Vol. 2, 39.


[25] Cited in [15], 295–6.

[11] Newton, I., *Philosophiae Naturalis Principia Mathematica. The Third Edition (1726) with Variant Readings* (ed. by Alexandre Koyré and I. Bernard Cohen, with the assistance of Anne Whitman). Cambridge University Press, Cambridge 1972. We follow the translation from the Latin provided in Isaac Newton, *The Principia: Mathematical Principles of Natural Philosophy.* A new translation by I. Bernard Cohen and Anne Whitman assisted by Julia Budenz, preceded by *A guide to Newton's Principia* by I. Bernard Cohen, University of California Press, Berkeley, Los Angeles, London 1999.

[12] Translated and cited in Shapiro, A., *Fits, Passions, and Paroxysms: Physics, Method, and Chemistry and Newton's Theories of Colored Bodies and Fits of Easy Reflection*. Cambridge University Press, Cambridge 1993, 25.

[13] Leibniz, G. W., *Leibnizens mathematische Schriften* (ed. by C. I. Gerhardt). 7 vols., Olms, Hildesheim 1971, Vol. 1 (2), 104. My translation.

[14] Niedersächsische Landesbibliothek (Hannover) Lh 35 VIII 21, f. 1r. Quoted in E. Pasini *Il Reale e l'Immaginario: la Fondazione del Calcolo Infinitesimale nel Pensiero di Leibniz*. Sonda, Torino 1993, 149n. My translation.

[15] Hall, A. R. *Philosophers at War: the Quarrel between Newton and Leibniz*. Cambridge University Press, Cambridge 1980.

[16] Guicciardini, N. *Reading the Principia: the Debate on Newton's Mathematical Mathods for Natural Philosophy from 1687 to 1736*. Cambridge University Press, Cambridge 1999.

Dipartimento di Filosofia e Scienze Sociali, Università di Siena, via Roma, 47, 53100 Siena, Italy

E-mail: niccolo.guicciardini@fastwebnet.it

# e-learning mathematics*

Sebastià Xambó Descamps† (*moderator*)
Hyman Bass, Gilda Bolaños Evia, Ruedi Seiler‡, and
Mika Seppälä§ (*panelists*)

**Abstract.** In addition to the current state of knowledge about the learning of mathematics and its aims in today's society, the main purpose of this paper is discussing ways of improving the process of learning, and especially, in that regard, the role of e-learning technologies. We chart the situation of e-learning mathematics as of December, 2005, including distance-learning or open university courses, and then we consider a number of areas where e-learning is likely to develop. Finally, we assess the impact of e-learning on the role of the new educators in mathematics.

**Mathematics Subject Classification (2000).** Primary: 97-xx, 97Uxx; Secondary: 00-xx.

**Keywords.** Online material, distance learning, e-learning, metadata.

## Presentation

by *Sebastià Xambó Descamps*

Following a suggestion of the Executive Committee (EC) of ICM2006 that came forth in the Fall of 2004, this panel has been promoted by the Conference of Spanish Mathematics' Deans [1].

After having formally accepted the invitation on December 16, 2004, the CDM Executive Committee discussed possible topics, until "e-Learning Mathematics" (eLM) was chosen and approved by both the CDM and the EC of ICM2006. Names to be invited as panelists were also decided, and it is a great satisfaction, and an honour, to be able to say that all accepted. On behalf of the CDM, my sincerest thanks to all.

If e-learning is learning by means of systems built on current computer and communications technologies, then the main interest of eLM is on what advantages e-learning can offer in the case of mathematics.

The main reason for choosing eLM is that the accelerated evolution of the e-Learning field is having, and will most likely continue to have, a major worldwide impact

*Panel promoted by the Spanish Conference of Mathematics' Deans.

†Partially supported by the European e-Content project "Web Advanced Learning Technologies" (WebALT), Contract Number EDC-22253.

‡Thanks for the support by the Bundesminister für Bildung und Forschung.

§Partially supported by the European e-Content project "Web Advanced Learning Technologies" (WebALT), Contract Number EDC-22253.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

on many aspects of the teaching-learning systems, at all levels, while offering, at the same time, new opportunities to professional mathematicians and to existing or new institutions, as for example in life-long learning. It is thus a topic that should greatly interest not only mathematicians in all walks of life, but also academic and political authorities everywhere.

This is why we imagined that the panel could aim at describing the situation of eLM as of 2006, outlining the most likely trends of its evolution in the next few years, indicating what the strongest impacts (positive or negative) in the mathematics teaching-learning systems will be, and charting the sorts of opportunities that will arise.

We are of course aware that such aims can only be attained by the panel in very broad terms, although this should be enough to bring forward a generally useful picture. For those wanting to have more detailed views, the references provided by the panelists should be a valuable resource to continue a journey that by all evidence has no return. For example, the articles in the recent book [2] will quite likely be serviceable to a wide range of readers seeking to know more about e-learning in general.

Let me continue with a few general remarks on learning, teaching and e-learning.

Mathematics, or mathematics knowledge, is a vast universe (let me call it $M$). It has many smaller interelated universes, of which we have a dim glimpse in the standard classifications.

Because of the increasing number of research mathematicians, and the availability of ever more sophisticated computational and communication tools, $M$ has undergone an extraordinary growth, and all indications are that this trend will continue in the coming years. To a large extend this blooming is explained because $M$ is both a source of deep beauty and the only precision method we have for modelling the physical universe.

In any case, the number of university students required to take mathematics courses is globally increasing, but at the same time the number of professional mathematicians that seek a teaching position is most likely decreasing, as there are, on one side, ever newer job profiles, and, on the other side, the number of students in mathematics degrees is decreasing in most countries. Moreover, in the last decade a steady decline in the mathematical skills of the students beginning higher education has been reported (see, for example, [3]).

Can eLM help to face this situation in a more positive mood?

The expectations created by e-learning are certainly high, at all levels, and we may wonder how much of it is going to be true, and up to what point can it help in the case of mathematics.

The reasons behind the high expectations on e-learning stem from well-known characteristics of the e-learning systems:

- In principle, access is possible from anywhere and at any time, thus making possible flexible (even just-for-me) and just-in-time courses of learning.

- The teacher can also be anywhere and do most of his teaching job at any time (preparing materials or following-up and coaching his students).
- It allows for synchronous activities of a teacher and a group (at an agreed time), but again without restriction on the location of the people involved, and, what is more, with the possibility of addressing a much larger audience than a conventional class.
- Assessment can be automated to a large extend and final grading can be integrated seamlessly into the institution's information system.
- The learning materials and experiences can be richer in many ways, and they can be easily maintained and updated (as compared to preparing, say, a new edition of a paper book).
- There are also indications that it may induce deeper understanding and stronger retention.

So the main question is how can we harness all that potential for improving the quantity and quality of the learning of mathematics. Since there are many levels that we ought to consider, and many variations in each level, we cannot expect a universal recipe. And even if we restrict ourselves to a very particular situation, say remedial mathematics for freshman in engineering schools or mathematical modules for prospective secondary school teachers, we cannot expect a formula that would satisfy everybody.

A sensible starting point is just looking at people, groups and institutions that are leading the way in one direction or another. This is the idea behind the purpose and composition of this panel. Since it is not feasible, and perhaps not even desirable, under the circumstances, to have a comprehensive survey of eLM, the best alternative is having experts in a few areas that have a major bearing on what eLM is and can be, and on how it is evolving. Before going into their reports, let me briefly introduce them.

**Hyman Bass**

Hyman Bass is Roger Lyndon Collegiate Professor of Mathematics and Professor of Mathematics Education at the University of Michigan. A graduate of Princeton, Dr. Bass earned his Ph.D. from the University of Chicago under Irving Kaplansky. He has had visiting appointments at sixteen different universities in ten countries. The many honors and prizes that Dr. Bass has received include the Cole Prize in algebra. He is an elected member of the American Academy of Arts and Sciences and the National Academy of Arts and Sciences, and the Third World Academy of Sciences, and was elected Fellow of the American Association for the Advancement of Science. He is former president of the American Mathematical Society and current president of ICMI. He has been both a Sloan and Guggenheim Fellow. Dr. Bass has published eighty-six papers in mathematics and seventeen in mathematics education.

**Gilda Bolaños**

Dr. Bolaños is a certified teacher and trainer in the didactical techniques of Problem Based Learning (PBL) and Project Oriented Learning (POL). She is the author of several certified Blackboard courses. With classroom technologies based on Maple and Minitab, she has worked extensively on problems and materials for her online courses.

**Ruedi Seiler**

Full Professor for Mathematics at the Technische Universität Berlin, Ruedi Seiler's main fields of interest are Mathematical Physics, Quantum-Hall Systems, Information Theory, Data Compression, and E-Math: Teaching, Learning, Research. Member of the Research Center "Mathematics for Key Technologies" (DFG), and of the Executive Committee of the International Association of Mathematical Physics (IAMP), his most recent undertakings, culminating an extensive experience in organizing events and participating in projects, are MUMIE and MOSES. More specifically, he is leading, since 2001, the project "Multimedial Mathematical Education for Engineers", a project developed in Cooperation between the Berlin University of Technology, the Munich University of Technology, the Aachen University of Technology and the University Potsdam (funded by the German Federal Ministry of Education and Research within the programme "New Media in Education"), and, within the program "Notebook-University" of the German Federal Ministry of Education and Research, he is co-manager, since 2002, of the TU Berlin project "MOSES – Mobile Service for Students".

**Mika Seppälä**

Dr. Seppälä is Professor of Mathematics at Florida State University and Professor of Computer Aided Mathematics at the University of Helsinki. He was the Co-ordinator of the HCM network "Editing and Computing" (1995–1996) which initiated the development that lead to the MathML and OpenMath languages allowing the inclusion of mathematical formulae on the web pages in a meaningful way. He is currently the Secretary of the OpenMath Society, and the co-ordinator of the eContent Project "Web Advanced Learning Technologies" (WebALT). The main goal of the WebALT Project is to use MathML and OpenMath to create tools and content for multilingual on-line mathematics. Seppälä was the President of the Finnish Mathematical Society for the period 1992–1996.

**Sebastià Xambó Descamps**

Full Professor of Information and Coding Theory at the Universitat Politècnica de Catalunya (UPC, Barcelona, Spain), and former Full Professor of Algebra at the Departamento de Algebra of the Universidad Complutense of Madrid (1989–1993), is serving as Dean of the "Facultat de Matemàtiques i Estadística" of the UPC. Member

of the EU eContent Project "Web Advanced Learning Technologies". In the period 1994-2000 led the team that developed the mathematical engine of Wiris ([4], [5]) and authored the e-book [6]. Cofounder of Maths for More ([7]). Has served as President of the Societat Catalana de Matemàtiques (1995–2002) and of the Executive Committee of the 3rd European Congress of Mathematics (Barcelona, 2000), and as Vicerector of Information and Documentation Systems of the UPC (1998–2002). Since the Fall of 2004 he serves as President of the Spanish Conference of Mathematics' Deans.

## References

[1] http://www.usc.es/mate/cdm/.

[2] *Integrated E-learning. Implications for Pedagogy, Technology and Organization*. Edited by Wim Jochems, Jeroen van Merriënboer and Rob Koper. RoutledgeFalmer, 2004.

[3] http://www.engsc.ac.uk/index.asp.

[4] Eixarch, R., Marquès, D., Xambó, S., WIRIS: An Internet platform for the teaching and learning of mathematics in large educational communities. *Contributions to Sience* **2** (2) (2002), 269–276.

[5] Eixarch, R., Marquès, D., Xambó, S., Report on the positive effects for an educational community of having universal Internet access to a mathematical computational system. In preparation.

[6] Xambó, S., *Block Error-Correcting Codes: A Computational Primer*. Springer-Verlag, 2003. Digital version at http://www.wiris.com/cc/.

[7] http://www.mathsformore.com/.

## The instructional potential of digital technologies

by *Hyman Bass*

**Educational uses of technology.** Digital technology continues to rapidly transform all aspects of life and work, even (and perhaps all the more so) in the developing world. It is designed, and presumed, to bring great benefit and empowerment to its users, as well as profit to its developers. Yet, as it opens new and even unanticipated possibilities, it poses as many problems as it solves, some new, and some techno-versions of classical problems, all of them important and interesting. And technology, for its novelty and glamorous aspirations, is greedy for our attention, liking to take center stage in every arena it enters.

Education, and mathematics education in particular, is the context in which this panel is examining these transformations. I find it helpful here to distinguish three broad kinds of roles that technology can play in mathematics education. They are of course not disjoint.

I. ***Transmission:*** Use of technology (web, video conferencing, etc.) to transmit, perhaps interactively, instruction and/or instructional materials that are conceptually of a traditional genre – lectures, demonstrations, problem sets, assessments, etc. These are the kinds of uses that fundamentally support distance learning, for example.

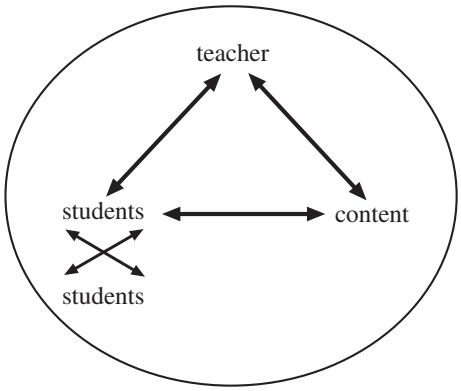
II. ***Power, speed, and visualization:*** Use of technology to carry out quickly and more accurately and completely, mathematical processes of a traditional nature – perform large or complex calculations, solve equations, approximate integrals, exhibit function graphs, study effects of variation of parameters, produce vivid and accurate images of geometric figures, etc.

III. ***New ways to explore the*** (***mathematical and experiential***) ***universes:*** Use of technology to do things we have never previously been able to do. Such capability affects mathematics itself, not just mathematics education. Examples include the study of long-term evolution of dynamical systems, and the images of fractal geometry that emerge there from. (This had an effect on dynamics comparable with that of the telescope in astronomy and the microscope in biology.) Software development gave life to the field of computational complexity, with its applications to coding and cryptography. Mathematical modeling and computer simulation supports a virtually empirical study of physical systems and designs. Dynamic geometry offers unprecedented opportunities to visually explore and analyze geometric structures, and to produce evocative imagery of dimensions three and four (using time). Computer algebra systems furnish unprecedented resources for solving equations. Much of this new technological power is now within reach of many students, and this raises possibilities of thereby expanding the horizons of the mathematics curriculum.

At a pragmatic level, technology thus offers resources to address two fundamental challenges of contemporary education – distance and demographics. Distance because many learners in need are physically remote from the sources of quality instruction and materials. Gilda Bolaños offers us an excellent survey of diverse modes of distance learning formats. Demographics because class sizes, particularly in introductory level mathematics courses, are too large to afford adequate instructor attention to individual student learning. (Bounding class sizes is often done at the cost of using instructors of highly variable quality.) In this case, technology affords various interactive formats for student work and assessment. These include the "virtual laboratories" described by Ruedi Seiler, and the interactive online materials (lectures, automatically graded homework, etc.) discussed by Mika Seppälä.

But independently of these practical needs, technology also offers possibilities for improving mathematics instruction itself. And the fundamental questions about the quality of teaching and learning do not recede when the instruction is mediated by technology; they only change their form.

**Instruction.** By "instruction" I mean the dynamic interaction among teacher, content, and students. I rely here on the "instructional triangle" that Cohen and Ball use to depict the set of interactions that they call "instruction" (Cohen and Ball, 1999).




Viewed in this way, instruction can go wrong in some simple but profound ways, for its quality depends on the relations among all of these three elements. When they misconnect, students' opportunities for learning are impaired. For example, if a teacher is not able to make the content accessible to students, framing it in ways that are incomprehensible to them, the chances that they may misunderstand are great. If students' interpretations of a task are different from the teacher's or the textbook author's intentions, then their work may be misrouted or take the work in unhelpful directions.

It may seem slightly strange, in the context of this panel, to propose the above representation of instruction. For, if you think about it, most descriptions of instructional uses of technology appear to reside exclusively on the bottom edge of the instructional triangle, absent the teacher. A tacit premise of some of this thinking is that somehow, the technology, with its interactive features, actually substitutes for the teacher, or renders the teacher obsolete, except perhaps as a manager of the environment. The viability of this view is a deep and important question, one that I shall not enter here except to make a couple of observations. One is that, in the most successful models of distance learning, it was found to be essential to have a tutor or facilitator available at the remote sites of reception of the materials, to respond to the many questions and requests that students would have, and that were not adequately responded by the technology environment. In addition, it was found to be important to have real time online questioning of the primary source available at certain times. In other words, prepared and transmitted material alone no more teaches a learner than does a textbook, unmediated by a teacher. The other comment is that interactive technology formats can at best provide well-prepared instructional materials and tasks, and respond to the student productions and questions that the software developers have anticipated and for which they have programmed responses. There are many domains of procedural learning and performance where this can be somewhat successful, though

the software, no more than a skilled teacher, cannot completely predict and prepare for all of what students may come up with. Moreover this uncertainty is all the greater once one enters into territory that is less procedural and involves more conceptual reasoning and problem solving.

In what follows, I identify five persistent problems of mathematics instruction and discuss ways in which technology can be deployed to address these. How these are actually used, however, would affect the degree to which they were helpful, so for each case, I point out its possible pitfalls.

**1. Making mathematically accurate and pedagogically skillful diagrams.** One problem faced by mathematics teachers at all levels is how to make clear and accurate diagrams that make the essential mathematical ideas plain to learners, and how to do so in ways that are manipulable for mathematical reasoning. Doing this by hand is often no easy task, whether the sketch is of slices of an ellipsoid in calculus, or sixteenths of a rectangle in fifth grade. Mathematical accuracy is one dimension of the challenge; featuring is a second - that is, making the instructionally key features visible to learners. In addition, instructors must manage these challenges fluently, using class time effectively. An instructor who can make diagrams accurately and helpfully, but who must use 10 minutes of class time to do so, loses effectiveness. Diagrams are also used for a variety of purposes: explorationally, to investigate what happens if certain elements are allowed to vary, or presentationally, to demonstrate an idea, an explanation, or a solution. This means, sometimes, the need for dynamics - translations, rotations, rescaling, variation of parameters. Often diagrams must be made in ways that map clearly to algebraic or numerical representations. Drawing software, or other design tools, can help. Important is the capacity to produce carefully-scaled diagrams, with the capacity for color or shading, and to be able to move elements of a diagram. Its use must be fast and flexible, helpful both for carefully designed lectures and for improvisation on the fly, in response to a student's question. Such software or tools can provide significant support for the use of diagrams in class, by both students and instructor. Making such software accessible to students increases their capacity for individual explorations and preparation for contributions in class. Students can quickly put their diagrams up for others' inspection, or support a point in class, in ways that are difficult to do when students go to the board to generate representations by hand. Using software tools to support the visual dimensions of mathematical work in instruction can significantly alter a major dimension of instruction and do so in ways that are mathematically accurate, pedagogically useful, and sensitive to the real-time challenges of classroom instruction where class periods are finite and time is a critical resource.

Software tools to support the making of diagrams can create problems, too. For example, if the tools are rigid or interfere with the purposes for making diagrams, or cannot be manipulated as desired, the representations may not be as useful as needed. Another problem may be that the use of such tools inhibits students from developing personal skills of appraisal and construction. If the tools quickly make

correct diagrams, students may not develop a critical eye with which to inspect them. If they never have to make a diagram themselves, they may remain entirely dependent on the software and not develop independent capacities for drawing.

**2. Making records of class work and using them cumulatively across time.** A second pervasive problem of mathematics instruction can be seen in the overflowing blackboards full of work and the slippery sheets of transparencies filled with notation and sketches, generated in class, and that vanish into weak memory when class ends. The record of class work (not just text or prepared materials), whether lecture, discussion, or exploration, is an important product of instruction. Under ordinary circumstances, this product vanishes and is thus unavailable for study or future reference, use, or modification. So acute is this problem that, too often, even during a single class, such work is erased (in the case of chalkboards) or slid away (as in transparencies). The work of that single class period is weakened for not being able to secure its place in evolution of ideas in the course. Moreover it is not available for students who may have missed a class.

When the work done in class is created or preserved in digital form, an archive of the mathematical progress of the class can become a resource for ongoing learning. It can then be easily accessed and transmitted remotely to others. Doing it "live" in class requires skill and dexterity on the part of the instructor. Making records of classwork afterwards (i.e., photographing the board with a digital camera) is easier but possibly less manipulable for subsequent class work. Important, too, is that everyone who needs to access these records can work on a common platform or that the format will work reliably across platforms.

**3. Alignment between classes and textbook.** Instructors, perhaps in response to student ideas or productions, may choose to depart from the text - in topic treatment or sequencing, or even topic coverage, and in the design of student activities and tasks. If the instructor creates these variations and alternative paths in electronic form, then a new text is created based on the instructor's design. This affords students access to the substance and course of the lessons. This gives license to flexible and innovative instruction, by affording the means to do so without disadvantaging students through disconnection from a text to be perused and revisited over time.

**4. Ease of access to the instructor between classes.** In the developed world, it is hard to imagine university instructors who do not maintain email (and web) connection with their students. This has made much more fluent and elastic the traditional functions of "office hours." Most student questions can be handled expeditiously, in timely fashion (though asynchronously), by email (perhaps with attachments), thus greatly reducing the need for face-to-face meetings, with their scheduling difficulties. And, as with the discussion above, these exchanges can contribute significantly to the record of the student's work and progress. When appropriate, an exchange between one student and the instructor can easily be made available to other students, thus changing an individual "office hour" into a group discussion. Pitfalls can exist with electronic

communications, of course. Misunderstanding is frequent when communication is restricted to text, without gesture, intonation, and the ability to demonstrate or show.

**5. The repetitive nature of individual outside-of-class sessions.** One feature of traditional office hours, or help sessions, is that they tend to be repetitive, processing over and over again the same questions and assistance with each new student or group of students. When such assistance is administered electronically, and it is seen to be germane to the interests of the whole class, it is an easy matter to copy the whole class, or perhaps selected individuals, on such exchanges. This puts to collective profit the considerable instructional investment made in one student, or group of students, and everyone gains, not least the instructor. An important consideration here is sensitivity to privacy issues and confidentiality. In particular, making individual student communications requires prior consent.

**Conclusion.** Technology continues to transform all aspects of our lives and work. It is already difficult to imagine how we once functioned without email and the web. We are still at the early stages of trying to understand and design the best uses of technology for mathematics instruction. I have pointed to some promising uses of technology to address some endemic problems of even traditional instruction. I have also tried to signal that the fundamental problem of developing quality teaching does not disappear just because instruction is mediated in technological environments.

### References

Cohen, D. K., and Ball, D. L., *Instruction, capacity, and improvement*. (CPRE Research Report No. RR-043). University of Pennsylvania, Consortium for Policy Research in Education, Philadelphia, PA, 1999.

## Distance learning today

by *Gilda Bolaños Evia*

The definition of distance learning has been modified over time, and today we have a variety of definitions. We will adopt the definition of Greenberg, in [Greenberg98], where contemporary distance learning is defined as "a planned teaching/learning experience that uses a wide spectrum of technologies to reach learners at a distance and is designed to encourage learner interaction and certification of learning".

In this section we will discuss the effects of some of the technologies used in distance learning education on mathematics and its effects on student's knowledge.

**Video taped lectures.** Since the introduction of videos to instruct students on different areas, many studies have been conducted to determine the effectiveness of these

methods. Some examples are [Beare 89], [Moore 96], [Russell 97], and [Pflieger 61]. On all of these studies the conclusion is that there is no significant difference on the achievement of students on video classes and regular classes. A three-year study involving 200,000 students and 800 public schools states:

> "... whereas most comparisons showed no significant differences, 119 were significant in favor of TV-taught students, and 44 in favor of conventionally taught students." [Pflieger 61].

We have to observe that on these studies the quality of the taught material was the same for video students and traditional students. Due to the lack of availability of similar studies for Latin America, we asked some professors and authorities that have been part of the VIBAS (video high school system) about the effectiveness of the system. In general they think that there is a significant difference in favor of traditional education, but this difference is not because of the video system, but mainly because of quality of materials and lack of availability of tutors. Moore and Kearsky converge to the same opinion in [Moore 96]. They also estimate that the difference is bigger in mathematics and physics. Coordinators of mathematics departments in public universities in Guanajuato State, Mexico, have noticed that students coming from video systems have a higher probability to fail its first math courses. They argue that their math knowledge is lower compared with regular students.

In the opinion of these authorities video taped lectures will tend to disappear, but not in the near future, at least for underdeveloped countries, because it is one of the cheapest forms to deliver distance education. They will be replaced by technologies as videoconferences.

At some universities video taped lectures are used inside the classroom for very specific concepts within the syllabus to present an expert opinion. Teachers at the Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) highly recommend this instrument for advanced courses and also to present interesting and attractive applications on elementary courses.

**Video conference.** Video conference has been used within higher education for more than a decade. Video conferencing is highly used for teaching sessions, teachers training, seminars and research. At many universities video conference is used as a tool to bring into the class an international experimented and recognized teacher to a large number of students. From the experience at ITESM, has been determined that the success of a video conference class depends on such factors as:

a) Quality of sound, images and degree of interaction.

b) Compatibility of the equipment with ingoing and outgoing signal places.

c) Availability and quality of material presented in the video conference.

d) Quick response to students questions.

e) A tutor on the conference classroom. At ITESM and at the University of Salle Bajío, coordinators of the video conference programs have found that for subjects such as mathematics, statistics and classes with "heavy contents", the presence of a tutor capable of answering students questions regarding the content of the conference makes a significant difference to the students learning and grades.

f) Tutor-student oral communication is very important, because when listening to the student, the teacher might understand some questions better orally than using other types of methods like the internet. Especially in the case of mathematics and statistics, it is very hard for students to write down some of their doubts, and this may cause problems, like using the mathematical language improperly, or overcoming technological barriers that make an extremely difficult task to write down a mathematical sign in a computer.

According to the Faculty of Education at The University of Plymouth [Plymouth], the future of videoconferencing is to incorporate video conference into web based systems, so teachers and presenters can sit in their own office or in a nearby studio and present a 'live' lecture in front of a camera attached to a web server. Using a simple switching device and several cameras, the presenter can provide remote participants with graphics, whiteboard, flipchart and other visual aids as well as alternative views of the local classroom, lecture room, etc.

**Online courses.** By experience at ITESM, the first step to success for online mathematics and statistics courses it is to convince students about the feasibility of the project. At this institution, full online courses are offered just for graduate students. It is also very important to have a quick response to student's questions, so they feel that "there is someone supporting them on the other side of the line".

A second step is to make sure that students can manage technology properly and have all necessary means to remain on line and to send and download information, documents, graphics, etc.

On a study conducted by Karr, Weck, Sunal and Cook [Karr 2003] at the University of Alabama to analyze the effectiveness of online learning in a graduate engineering mathematics course, they divided the class into three groups: Group A (Online course only), Group B (traditional for the first two thirds of the course and traditional and online for the final third of the course, Group C (traditional on the first third, online for the second third and traditional and online for the final third of the course). On this study they found that:

a) Students perform better on the analytical portion of the course when they had used the online mode of delivery. According to the teachers and students feedback this is due to the consistency of online materials and the fact that they have to "face the problem on their own"

b) Students taking the class by traditional mode perform better on the in-class portions of examinations. This might have been because of the instructor

dropping inadvertently little hints about which aspects of the class might be on the test.

c) The two groups with a traditional mode segment perform better when they have access to both modes of deliver, traditional and online.

d) There was no significant difference on the overall performance of the groups.

From my personal experience and from some non formal studies conducted on high school and undergraduate courses it is reasonable to believe that similar results will be obtained for high school and undergraduate mathematics courses.

Many universities as ITESM consider, even for traditional courses, that online sections and online materials should be included to make courses more attractive to students and to enhance the student's performance, especially in traditionally difficult courses as mathematics.

**Online problems and materials.** Online problems are widely used to improve students learning on mathematics courses. Within the experience of Professor Maritza Sirvent and me, some advantages of using online problems on web based programs for mathematics problems are:

a) The bank of problems is large and includes a big variety of questions.

b) The students know immediately if their answer is correct, so they get engaged and they try the problem as many times as necessary to get the right answer.

c) Some students feel that using the computer helps them in their homework.

d) It is clear that the correctness of the problem is independent of the procedure used on the resolution. So they try their own ideas to solve the problem and use techniques as approximations using calculators. After that they study a method that will work at different situations.

e) Problems solved for students at the same class are similar but not the same so they can't copy the homework from a classmate.

f) Student's attitude toward mathematics problems seems to improve.

A disadvantage of online problems might be that, when entering the answer to a problem, sometimes the student makes a typing mistake or forgets some parenthesis and then gets an incorrect answer even if he has solved the problem correctly. Also students are not forced to write down the complete procedure, so when they are tested on a traditional writing test, they have no training on that.

WeBWorK is an internet based program to deliver homework to students on internet. It was designed by the University of Rochester. On a study conducted at Rutgers University to measure how effective WeBWorK was in improving learning measured by student's performance in Calculus [Weibel 2002], they divided students in two sections: Sections where WeBWorK homework was required weekly and it counts as part of the final grade, and sections where traditional written homework was required. Two thirds of calculus students were on WeBWorK sections, and they found the following:

a) Students in WeBWorK section did slightly better than students on traditional section. However, within WeBWorK sections, students who did over 80% of the WeBWorK problems performed dramatically better (by a full letter grade) than those who did less than half of the WeBWorK problems.

b) First year calculus students were very responsive to WeBWorK and most of them attempted every problem. They found that there is a 2-letter grade difference (on the average, from B to D) between students who do well on WeBWorK and those who do not attempt it. For upper class students taking calculus there is a 3-letter grade difference (on the average, from B to F) between students who do well on WeBWorK and those who do not attempt it. These upper class students are not very responsive to WeBWorK.

c) Students repeating calculus are not responsive to WeBWorK, and there is no significant difference on grades even for those that perform well on WeBWorK.

Online didactical material helps students to understand some concepts that might be difficult to them. Some students express that it is easier for them to read online materials than books, because they are usually more attractive and often interactive. For them it is the perfect complement for text books.

There is a bright future for online mathematics problems and didactical material. Each year the number of teachers convinced of the effectiveness of online mathematics problems and didactical material is increasing. Internet-based methods to deliver homework to students are improving and making it easier for teachers and students, saving a considerately amount of time on grading. For instance, projects such as WebALT [WebALT] aim at using existing technology standards for representing mathematics on the web and existing linguistic technologies to produce not just online mathematics problems, but language-independent mathematical didactical material.

**Problem based learning (PBL) and project oriented learning (POL).** These learning methodologies have been applied from elementary school to graduate programs. It is based on the principle that learning occurs not by absorbing information but by interpreting it. These methodologies are ideal for distance learning, but require that students work in teams, an arrangement that may be very difficult for some students that prefer to work individually. With these didactical techniques, learning is generated by solving a realistic situation that requires learning new concepts and applying them to solve a problem. At some universities the full curricula is build around PBL or POL techniques, while at some other universities (as ITESM) these methodologies are mixed with traditional methods [Bolaños 2003], [Watson 2002]. PBL and POL are excellent tools to introduce students on the more difficult tasks of the syllabus. The results are excellent, as statistics show that students perform better with the concepts when introduced by PBL or POL than when introduced on traditional lectures.

On these methodologies the role of the tutor is very important. The tutor is responsible for the direction of students and to help in team conflicts. The tutor has to

address the student's efforts in the right direction and make suggestions about working lines. Students communicate online with their teammates and the tutor, also the final report of all teams is placed online, and so all teams might look at the similarities and differences with the solutions of the others teams.

These are just some aspects of the big world of distance learning and were choosen because we consider that they might be applied on very different teaching environments. Distance learning will continue modifying our teaching practices.

## References


[Beare 89] Beare, P. L., The Comparative Effectiveness of Videotape, Audiotape, and Telelectures in Delivering Continuing Teacher Education. *The American Journal of Distance Education* **3** (2) (1989), 57–66.

[Bolaños 2003] Bolaños, G., Problem Based Learning for great statistics learning. In *Proceedings of the Hawaii International Conference on Statistics and Related Fields*, 2003.

[Greenberg 98] Greenberg, G., Distance education technologies: Best practices for K-12 settings. *IEEE Technology and Society Magazine* (Winter) 36–40.

[Karr 2003] Karr, C., Weck, B., Sunal, D. W., and Cook, T. B., Analysis of the Effectiveness of Online Learning in a Graduate Engineering Math Course. *The Journal of Interactive Online Learning* **1** (3), 2003.

[Moore 96] Moore, M., and Kearsley, G., *Distance Education: A Systems View*. Wadsworth Publishing Company, Belmont 1996.

[Moore 97] Moore, M., and Thompson, M., The Effects of Distance Learning, Revised Edition. Technical Report ACSDE Research Monograph (Number 15), American Center for the Study of Distance Education, The Pennsylvania State University, 110 Rackley Building, University Park, PA 16802-3202, 1997.

[Pflieger 61] Pflieger, E. F., and Kelly, F. G., The National Program in the Use of Television in the Public Schools. Technical Report, Ford Foundation/FAE, 1961.

[Plymouth] University of Plymouth web page, http://www2.plymouth.ac.uk/distancelearning/vidconf.html

[Russell 97] Russell, T., The "No Significant Difference" Phenomenon. Retrieved from http://tenb.mta.ca/phenom/phenom.html, 1997.

[Watson 2002] Watson, G., "Using technology to promote Success in PBL Courses". The Technology Source, 2002.

[WebALT] WebALT web page, http://webalt.math.helsinki.fi/content/index_eng.html.

[Weibel 2002] Weibel, C., and Hirsch, L., WebWork Effectiveness in Rutgers Calculus. Retrieved October 27, 2005, from http://math.rutgers.edu/∼weibel/ww.html, 2002.

[Whittington 89] Whittington, N., Is Instructional Television Educationally Effective? A Research Review. Readings in Principles of Distance Education. Pennsylvania State University, 1989.

# Virtual labs in mathematics education: concepts and deployment

by *Ruedi Seiler*[1]

**Background.** The work field of engineers, as well as that of scientists and mathematicians, is undergoing drastical changes: as numeric software and computer-algebra-systems are capable of performing intensive and complex arithmetical calculations, other abilities, such as the fast acquisition of new knowledge and new methodologies, are growing in significance. Thus, learning and teaching methods that promote life-long, efficient and independent learning have to be conveyed.

The traditional teaching methods employed at universities are of only limited success in this respect: Teacher-centered lessons provide the essential basic knowledge, but it hardly allows for a more active approach to the subject-matter. Classical experiments, in contrast, while targeted at independent knowledge acquisition, soon stumble across limits imposed by the reality of a university: high and constantly increasing numbers of participants in a course, limited access to and inadequate equipment. In addition, the experimental approach to knowledge acquisition in "real laboratories" is by its very nature limited to certain fields of studies, while more theoretical fields, such as mathematics and theoretical physics are either completely precluded or only peripherally touched upon by the existing experimental concepts.

The deployment of new media and technology in class thus represents a turning point: *Virtual Labs* are environments based on physical labs in which computer aided experiments can be designed, created, implemented and evaluated. Experiments are implemented in the form of computer-based algorithms, representing either real tools and objects or even theoretical concepts and objects.

Such explorative learning environments can be placed at the disposal of every student and teacher, independent of time and place. In the framework of the classical experimental sciences, virtual labs are capable of complementing real laboratories by allowing the concise elaboration of the actual "phenomenon" and diminishing the influence of metrological problems. As, however, the handling of the equipment and the mentioned problems represent a vital part of the acquired competence, real laboratory experiments should not be set aside completely in the experimental disciplines. In theoretical subjects, on the other hand, these technologies make abstract phenomena visually comprehensible.

In this article, we will offer detailed requirements on Virtual Labs and describe the consequences of the implementation along the lines of a prototypical Virtual Lab for Statistical Mechanics.

**Pedagogical requirements.** In the following, we present a list of pedagogical requirements we demand from modern e-learning technology, especially from virtual

[1] In collaboration with Thomas Richter (TU, Berlin, thor@math.tu-berlin.de) and Sabina Jeschke (TU, Berlin, sabina@math.tu-berlin.de).

laboratories. In comparison with most other e-learning environments, though, virtual labs do not define learning goals by themselves. Rather, they put "learning spaces" at the disposal of teachers and students.

A laboratory should provide the necessary equipment – or, in the case of virtual labs, the necessary algorithms – that facilitate the independent development and testing of *problem solving strategies*, incorporating typical problems of mathematics, physics and engineering science in order to prepare the student for his or her professional life.

Laboratories offer students the unique opportunity to control their learning, without outside interference and consequently being able to make an independent decision about their learning process. We divide the support of self-directed learning into the following categories:

First of all, (Virtual) Laboratories support explorative learning by allowing their users to work independently and efficiently with the technical equipment in order to investigate interconnections independently and to build an intuitive understanding of the subject. Therefore, it is vital that Virtual Laboratories should allow and encourage unconventional approaches, options, work flows, etc.

Second, the support of different learning styles is one of the utmost features of the deployment of multimedia technologies in education, even though the first generation of e-learning technologies [1] did not yet allow individual approaches to the subject. Similarly, pre-fabricated experiments might not fit into the previous knowledge of the user, strictly limited specific environments and learning goals might not fit the individual interests, failing to motivate the user. Thus, virtual laboratories must enable the user to setup and control the experiment *freely*.

Laboratories should ideally be adaptable to different application scenarios. This includes the deployment of the same basic lab in different courses, stressing different field-specific foci on the one hand, and the use in different scenarios ranging from demonstration through practice to examination on the other hand. For that reason, a virtual laboratory should not be limited to a fixed set of experiments or aimed at the requirements of one single lecture or one specific audience; for each different target audience arise different requirements. Typical application scenarios might reach from simple demonstrational support within lectures, over experiments in the classroom teaching for training and tutorials up to self-study and deployment in research applications.

Both research and engineering achievements are increasingly the result of cooperations between distributed, separated teams. Thus, team work and team-oriented projects have to be an integral part of any modern scientific education, and thus must be actively supported by virtual laboratories as well.

Laboratories must offer appropriate interfaces that will allow the integration of or linking with standard elements as Maple or Mathematica; experimental set-ups should include these elements correspondingly, for their use and handling should be a part of the scope of learning.

Laboratory elements should be detachable from the actual lab through the application of open interfaces and thus should be reusable. Such requirements not only

allow the efficient construction of new laboratories from existing elements, they also ease the integration of laboratories in more complex experiments requiring additional support from outside software components.

**Consequences for the implementation.** The pedagogical requirements on virtual laboratories pose various demands on the software design which we demonstrate for the laboratory VideoEasel developed at the DFG Research Center. The technical focus of this laboratory is in a first, prototype phase with application to the field of statistical mechanics and related areas. Statistical problems are here modeled through the use of cellular automata, which are well-suited to design statistical models, covering many interesting areas ranging from the Ising model, statistical image denoising, lattice-gas models to Turing completeness.

In order to be able to support different and varying deployment scenarios while imposing as few restrictions on the labs themselves, it must be possible to combine the elements of laboratory equipment flexibly and creatively. This leads to a "strictly anti-monolithic", fine-granular software design, its basic structure characterized by the tripartition into simulation and arithmetic modules implementing the mathematical modules, an interface layer that serves as link between the equimentments, that allows the free combination of the software modules into an experiment, and last, graphical user interfaces allowing to control the experimental setup conveniently.

The experiments in the lab VideoEasel are implemented as small, modular units, independent of the lab's actual core, that can be created and loaded on demand. The elementary units can be separated into two distinct classes, "automata" for the algorithmic definition of physical phenomena – e.g. the Ising model – and "measuring tools" to measure certain quantities arising within the experiment – e.g. the Free Energy. VideoEasel offers basic methods for evaluation of measurements, but does not provide any numerical tools for more complex analysis or a build-in process control for more elaborate experiments. Such functions are taken over to specialized tools by utilizing the software interfaces of the laboratory, which are here realized in the middle-ware CORBA [2]. Mappings are available to many languages, such as Java, C and Python, thus facilitating the connection to various other external tools. Presently, in addition to the native Java-interfaces, there is a Python-connection for script-control, as well as a C-implementation of a Maple-connection available.

Cooperative learning strategies in virtual laboratories imply in particular that several users from different working locations can work simultaneously on a single experiment while being well aware of the actions of their partners. Therefore, the need of designing the laboratory as a multi-part network application becomes self-evident: experiments are, for example, run on a server accessible by students.

VideoEasel follows a classical client-server approach where the students control the simulations run on the server by Java front-ends. In the most simple case – as for support of a lecture in a auditorium – server and client are run on the same computer; in cooperative learning settings, the server synchronizes more clients.

Reading the above arguments concerning the requirements in implementing a virtual laboratories drafted in the previous paragraph might create the impression of a "canonical" approach. However, most existing virtual laboratories posses a narrow technical focus on specific areas and follow a monolithic design.

The second remark concerns tutorials, user guidance and the "usability" of such laboratories: The afore-mentioned flexible granular structure of the software inevitably leads to a more complex user interface and consequently to a higher adaptation time for the teaching staff as well as the students. Problems arising from the initial contact with technical problems present a prominent "motivation killer" in e-learning. In some cases, it is not easy to find the ideal compromises; to overcome this problem, one should then provide several, separate user interfaces, as for example found in VideoEasel:

For simple demonstrational applications in lectures, a Java Applet is available that allows only minimal control of an experiment. For deployment in student groups and classroom teaching, a simple but efficient Java interface has been developed; it provides more options to influence the experiment, while keeping the complexity rather low. Additional menus allow the adjustment of all kinds of parameters within the experiment. The drawing surface, though, is very similar to the applet and mimics that of standard software tools.

A more refined and complete interface was created through the Oorange toolkit [3] – also developed at the TU Berlin – allowing the purely graphic set-up of an experiment, as well as the integration and connection to other elements through "Java Beans" [4]. The server provides templates available for existing experiments, similar to the ones for the Java interfaces; these templates are transformed client-side into a Oorange compatible XML-representation. Different from the more basic interfaces, the user has the option of changing, modifying or completing the experiment at will. This access to VideoEasel does not have the pretense of being particularly easy to navigate, as it was conceived primarily for the use in research and not in teaching or in practice. Therefore, it is acceptable to require the user to go through a reasonable adaptation phase.

Last but not least, VideoEasel is also completely controlable from within the computer algebra program Maple for applications whenever the Oorange toolkit is not able to deliver the mathematical algorithms required for research purposes. This interface uses, similar to all others, the CORBA technology to exchange data between the components.

Now, in retrospective, we analyze how the required didactic concepts are implemented within VideoEasel: the field of cellular automata is rich enough to simulate interesting physical effects, yet straightforward enough to avoid undue obstacles in easy access. The basic principle of such automata can be learned quickly and allows for the execution of interesting (and esthetically pleasing) experiments through quite basic tools. Through the integration of time-proven, well known concepts – drawing programs and measuring tools – and the choice of an appropriate interface, the user is encouraged to experiment. Comprehension of the behavior of the effect to be under-

stood is achieved through practice in the laboratory. Explorative learning is promoted through the connection of esthetical and academical contents.

The availability of various surfaces allows us to address several user groups with very different demands on the laboratories and diverse application purposes ranging from pure demonstration to research applications.

Cooperative deployment scenarios become viable through the two-part set-up as a client/server network architecture. Thus, acquisition and research between teams geographically far separated is feasible.

Finally, CORBA-interfaces allow the docking and linking of the core laboratory with other laboratories, algebra-systems and connectors to demonstrate even more complex facts and to avoid locking the user in one single laboratory technology.

**Future developments.** In conclusion, we will discuss some aspects of important relevance to our original aims, which are improving university education through the use of virtual laboratories:

Virtual labs, including the presented VideoEasel, are still mostly at a prototype stage. Thus, practical experience about their deployment in e-learning environments are still rare. It has to be expected that use and evaluation will result in extensive adaptations and expansions of the existing concepts, particularly in the field of usability.

To realize the pedagogical goals as presented above, it is necessary to integrate virtual laboratories into the framework of larger virtual knowledge spaces. VideoEasel does provide a number of generic interfaces which will have to be specified in more detail. More experiences with laboratories from other fields of science and engineering are necessary to define a standardized data-exchange between different laboratories.

Finally, the virtual laboratories are becoming more and more complex to use as a direct result of the diversity of addressed learning scenarios, the desired interconnectability of different applications and the broad variety of the learning contents. To counter this effect it might be desirable to extend laboratories by "digital assistants" [5]. New concepts developed in the field of artificial intelligence in recent years have to be expanded and applied to virtual knowledge spaces and their components.

## References

[1] Jeschke, S., and Kohlhase, M., and Seiler, R., eLearning-, eTeaching- & eResearch-Technologien - Chancen und Potentiale für die Mathematik. *DMV-Nachrichten*, July 2004.

[2] Scallan, T., A Corba Primer. http://www.omg.org/.

[3] Oorange: The Oorange development environment. http://www.oorange.de/.

[4] *JavaBeans*. http://java.sun.com/products/ejb/.

[5] Jeschke, S., and Richter, T., and Seiler, R., Mathematics in Virtual Knowledge Spaces: User Adaption by Intelligent Assistents. In *Proceedings of the 3rd International Conference on Multimedia and ICTs in Education*, June 7–10, 2005.

## Roles for the new mathematics educators

by *Mika Seppälä*

**The future is here. It is just not evenly distributed.** We are living interesting times! The industrial revolution is on its way in education, publishing, and business. Ways to conserve knowledge and transfer it from generation to generation are changing. Libraries are becoming digital and classes virtual. This development opens extraordinary opportunities to those willing and capable to profit from them. It also opens possibilities to spectacular failures of which we saw many some years ago.

"Emergent technology is, by its very nature, out of control, and leads to unpredictable outcomes." This certainly applies to the current development in e-learning, including e-learning mathematics. "The Future is here. It is just not evenly distributed." Both quotes are by William Gibson.

So in order to understand what lies in the future we can simply look at what our colleagues are doing today. There is no doubt that the information network and the advanced technology are going to change the way we write, publish and teach all disciplines, including mathematics, in the future. This will happen because it is possible, and because proper usage of technology will enhance our current ways to work.

To understand how educators work in 2016, we simply need to understand which, of the currently existing ways to use information technology in education, have most potential. These are likely to emerge as general paradigms and set examples that many will follow.

**Changing the educational system.** Not only instruction, but the whole educational system is changing. New interdisciplinary fields are emerging at a fast pace. Largely this is due to mathematics becoming more applicable thanks to the various advanced mathematics systems like Maple, Mathematica or Matlab. It is now possible to use mathematical modeling in a fundamentally deeper way than before. This is true in practically all fields, perhaps most notably in biology and medicine.

In the past, applications of mathematics in biology or medicine have been, from the mathematical point of view, rather simple. Now more complex methods can be used. This requires expertise in mathematics, computer science, and in the subject matter to which mathematics is being applied. Hence interdisciplinary study programs have been created to educate experts capable of developing these new applications.

**The new roles of mathematics educators.** In the past, and in many cases even today, the teaching of mathematics has been the responsibility of instructors, and the learning that of students. At most European universities, basic mathematics courses are being taught in very large sections. A typical undergraduate calculus class may have well over 100 students. In some cases these classes have hundreds of students.

The instruction is lecturing with little or no personal interactive contact between the students and the professor. Instructors simply cannot follow the day-to-day progress of their students.

Technology can be very useful here. Using systems like Maple TA or STACK, it is possible to offer automated private instruction to students and to monitor the progress of individual students even in large classes. This will empower professors and enhance traditional contact instruction in a dramatic way.

Instruction, even in the case of large classes, becomes student centered instead of instructor centered. Professors will take responsibility of their students in a way that has not been usual in the past. The emerging new role of instructors is very similar to that of coaches. Athletes have their personal coaches, so will students as well. The future instructors work like sports coaches today assisting students to achieve goals they could not achieve on their own. Empowered with advanced learning technologies, instructors can provide individual assistance to their students in a way that was not possible earlier. Interactivity can now be provided, using the web, in a way that is likely to permanently change the way we work.

**Educating new educators.** The inertia of the academia resists changes and delays the necessary development. Instructors in general are not ready to change the way they work. There is also a good reason for the resistance. Moving from traditional contact instruction to computer aided learning is not easy. The data in the table below are generally accepted estimates of the efforts needed for various types of teaching.

All these forms of teaching, except lecturing and small group teaching, will require additional technical support. The large spreads in the first four items reflect the fact that experienced educators can work much faster than beginning professors. There is no spread in the table for computer aided learning and interactive video. Here also experience will eventually help, but for now there are not many instructors having extensive experience in computer aided learning.

**Academic work to produce one hour of student learning** ([2])

| | |
|---|---|
| Lecturing | 2–10 |
| Small group teaching | 1–10 |
| Videotaped lectures | 3–10 |
| Authoring a text | 50–100 |
| Computer aided learning | 200 |
| Interactive video | 300 |

Using the figures of the above table, the development of a typical one semester course will amount to over five years of full time work of the author in addition to the required technical support.

Regardless of the above, some professors are developing content for computer aided learning. They are driven by the vision of greatly improved education once

the necessary content is in place and available in the same ways as books are now available to students and professors.

**Metadata.** Developing content for computer aided learning is very costly. Furthermore, today the materials developed by professors are mostly being used only by the authors themselves and their students. Sharing does not happen, not to speak of shared development of content. To address this problem, the European Commission is currently investing heavily into projects which enhance existing content with metadata. This metadata will make content cross border usable, and shared creation of content a real possibility. The development of metadata is likely to dramatically change the way we work. It will make the hard work to develop premium on-line content cost effective and worthwhile.

**Best practices.** All of the above applies really to all disciplines. Problems related to the teaching of mathematics or of sciences more generally, focus on the presentation of mathematical formulae. Almost all mathematics is being written using LaTeX today. Also these proceedings have been prepared by LaTeX.

LaTeX and TeX generate extremely high quality typesetting of scientific text. These systems produce content ready for printing and publishing in the traditional way. New LaTeX classes for producing high quality presentations have been created. Practically all mathematicians are using LaTeX.

Intelligent interactivity ([3] and [4]) requires that mathematical formulae are presented in the on-line materials so that the meaning of the formulae can be automatically understood. MathML and OpenMath make this possible. To embed mathematical formulae in a proper way to web content requires the usage of these languages. LaTeX or TeX do not support MathML or OpenMath. In spite of the fact that TeX enthusiasts are working hard to develop solutions to this problem, the use of MS Word and PowerPoint together with products like MathType often makes the content development much easier.

Searching the web one can find, for example, a variety of electronic presentations of calculus or linear algebra courses. Most of these are pdf presentations of printed materials, and are not designed to be studied from the computer monitor. The new media, the computer screen, requires a different presentation of the content than what is used on printed materials. The resolution of a printed page is much higher than the resolution of the best monitors. Hence printed pages are easier to read than computer monitors. To overcome this problem, content, for the computer screen, needs to be presented in a very condensed way. For instruction based on the computer screen, the presentation of the materials needs to follow the general design principles implemented, for example, in PowerPoint.

On-line content has many important advantages which greatly overcome the handicap that computer monitors have with respect to printed pages. These advantages include hyperlinking, live interactive and adaptive content, student performance track-

ing, and, most recently, multilinguality. The WebALT encoding of mathematical content uses an extension of OpenMath and is such that the content can be generated in many languages automatically. Hence the content is truly multilingual, or rather, language independent. This is a serious advantage in view of the high cost of the development of on-line content.

**A case study: on-line calculus at the University of Helsinki.** The lesson learned from previous experiences at Florida State University was that on-line materials should use standard tools as much as possible, not require students to install new programs, and that the illustrations of mathematics should be done so that the required technicalities are completely hidden.

With these points in mind, the development of new on-line materials for calculus was started at the University of Helsinki in the Fall of 2001. These materials consist of a collection of lectures presented by PowerPoint, a collection of PowerPoint presentations of solved problems, a collection of calculus calculators empowered by MapleNET, and a repository of problems delivered to students using Maple TA, a system for the delivery and automatic grading of homework, quizzes, and examinations.

Students reactions to these new on-line materials have been overwhelmingly positive. During the Fall of 2004, a basic course in calculus was offered, at the same time, as a fully on-line course, and as a traditional lecture/problem session course. Both courses were based on the on-line materials, and had the same exercises and examinations. For the on-line students, the examinations were the only events that took place on campus and were proctored.

The results were surprising: the on-line students fared better than the traditional students in both examinations, and the retention rate was higher among the on-line students than among the traditional students.

**Automatic assessment.** Systems providing automatic assessment of homework problems, quizzes and examinations have been used in lower level mathematics instruction at Florida State University with spectacular results for several years. The failure rates of precalculus courses have gone down by about 50%. This is due to students being able to practice for examinations at home so that they get immediate feed-back from the system.

Currently the most advanced automatic assessment systems are Maple TA, STACK, the forthcoming LeActive Math System and the WebALT System. Common to all of these is that they offer the possibility to create algorithmic problems which are programs that generate a different version of a problem every time the program is invoked. In addition to the others, the WebALT System will also be able to generate the problem in many languages.

The algorithmic problems really make a difference. Consider, for example, the method of partial fraction decompositions. Students of calculus will have to learn that. It is relatively simple to write a program which generates over a million different but

equally hard problems of partial fraction decompositions. Hence the examination about partial fraction decompositions can be published to the students before the test! Students can take the partial fraction decomposition test as many times as they want at home, get individual feed-back including full solutions. Learning by heart is not helpful, because regardless how many times they take the same test at home, they are going to get different questions in the examinations.

Such algorithmic problems were used in instruction at Florida State University in Spring 2005. Most students reacted very positively, and used the system a lot to their benefit. Some students solved even hundreds of problems on computing limits, for example. Starting in Fall 2006, students at Florida State University are required to have a laptop computer. Then the automatic assessment systems can be used in class, and examinations can be based on the use of these systems.

**Conclusions.** The development on-line education in mathematics at the university level has been very slow. Administrators at national agencies and ministries in various countries see the great potential that on-line content can bring to education, but largely this potential has not been realized in mathematics and, more generally, in sciences.

This is partly due to problems that one has in the presentation of scientific content on the web. The majority of on-line materials present mathematical formulae as pictures only. This is not a satisfactory solution. One cannot use a picture as a key word in a database search.

MathML and OpenMath provide solutions to this. Commercial editors, such as MS Word and PowerPoint together with MathType, provide a convenient way to produce content in which mathematics is embedded using MathML. Authoring tools are available, robust and easy to use.

Missing synchronous interactivity has been another problem in on-line instruction. Together with the introduction of tools like Skype and the various easy-to-use conferencing systems, this problem has suddenly disappeared. Virtual on-line courses can provide more personal interaction between instructors and students than a regular class with hundreds of students attending the same lectures. This development is new, and we have not yet seen how that will change instruction. The effect is likely to be impressive, however.

To use the available technology to the maximum places large demands on instructors. They have to rethink their roles and convert themselves from lecturers to coaches. And they have to be able to use technology in a fluent way. Most instructors resist doing this mainly because the transition requires a lot of work.

The main remaining obstacle in this development is the fact that premium on-line content is expensive to produce and hard to find. Extensive funding programs, like the European Commission supported Content Enhancement Projects of eContent Plus, are likely to make a dramatic difference with respect to these remaining obstacles.

The most important lessons learned were that it is necessary to keep the use of technology as simple as possible while still providing advanced functionalities. Pretty

good is good enough. For the student, everything has to work right out of the box. Technicalities have to be hidden. On-line content satisfying this criteria is going to have large and permanent value. In 2016 we cannot understand how education without the information network and its services was possible.

## References

[1] Bass, H., Mathematics, mathematicians, and mathematics education, *Bull. Amer. Math. Soc.* **42** (2005), 417–430.

[2] Boettcher, Judith V., Designing for Learning. http://www.designingforlearning.info/services/writing/dlmay.htm.

[3] Caprotti, O., Seppälä, M., Xambó, S., Using Web Technologies to Teach Mathematics. In *Proceedings of SITE 2006 Conference*, Orlando, FL, March 20–24, 2006.

[4] Caprotti, O., Seppälä, M., Xambó, S., Mathematical Interactive Content: What, Why and How. To appear in *Proceedings of the 1st WebALT Conference* (held in the Technical University Eindhoven, January 5–6, 2006).

[5] Grottke, S., Jeschke, S., Natho, N., Rittau, S., Seiler, R., **mArachna**: Automated Creation of Knowledge Representations for Mathematics. To appear in *Proceedings of the 1st WebALT Conference* (held in the Technical University Eindhoven, January 5-6, 2006).

Facultat de Matemàtiques i Estadística, Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: sebastia.xambo@upc.edu

School of Education, University of Michigan, Ann Arbor, Michigan, U.S.A.
E-mail: hybass@umich.edu

Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, México
E-mail: gbolanos@itesm.mx

Institut für Mathematik, Technische Universität Berlin, Berlin, Germany
E-mail: seiler@math.tu-berlin.de

Department of Mathematics and Statistics, University of Helsinki
and
Department of Mathematics, Florida State University
Helsinki, Finland & Talahassee, USA
E-mail: mika.seppala@webalt.net

# Author index